Variable Selection in Varying Coefficient Models for Mapping Quantitative Trait Loci

Yi Gong

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2011

Approved by:

Fei Zou, Advisor

Fred Wright, Committee Member

Haibo Zhou, Committee Member

Wei Sun, Committee Member

Tim Wiltshire, Committee Member

Abstract

YI GONG: Variable Selection in Varying Coefficient Models for Mapping Quantitative Trait Loci.
(Under the direction of Fei Zou.)

The Collaborative Cross (CC), a renewable mouse resource that mimics the genetic diversity in humans, provides great data sources for mapping Quantitative Trait Loci (QTL). The recombinant inbred intercrosses (RIX) generated from CC recombinant inbred (RI) lines have several attractive features and can be produced repeatedly. Many quantitative traits are inherently complex and change with other covariates. To map such complex traits, phenotypes are measured across multiple values of covariates on each subject. In the first topic, we propose a more flexible nonparametric varying coefficient QTL mapping method for RIX data. This model lets the QTL effects evolve with certain covariates, and naturally extends classical parametric QTL mapping methods. Simulation results indicate that the varying coefficient QTL mapping has substantially higher power and higher mapping precision compared to parametric models when the assumption of constant genetic effects fails. We model the time-varying genetic effects with functional approximation using B-spline basis. We apply a nested permutation method to obtain threshold values for QTL detection. In the second topic, we extend the single marker QTL mapping to multiple QTL mapping. We treat multiple QTL mapping as a model/variable selection problem and propose a penalized mixed effects model. We apply a penalty function for the group selection of coefficients associated with each gene. We propose new selection procedures for tuning parameters. Simulations showed that the new mapping method performs better than the single marker analysis when multiple QTL exist. Last, in the third topic, we extend the multiple QTL

mapping method to longitudinal data. We pay special attention to modeling the covariance structure of repeated measurements. Popular stationary assumptions on variance and covariance structures may not be realistic for many longitudinal traits. The structured antedependence (SAD) model is a parsimonious covariance model that allows for both nonstationary variance and correlation. We propose a penalized likelihood method for multiple QTL mapping using the SAD model. Simulation results showed the model selection method outperforms the single marker analysis. Furthermore, the performance of multiple QTL mapping will be affected if the covariance model is misspecified.

Acknowledgments

I would like to express my deepest appreciation to my advisor Dr. Fei Zou for her continuous guidance, support and patience throughout my graduate study and thesis work. I also want to sincerely thank my committee members, Dr. Fred Wright, Dr. Haibo Zhou, Dr. Wei Sun and Dr. Tim Wiltshire for their constant encouragement and helpful suggestions.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 QTL mapping

1.1.1 Methods for QTL mapping

Quantitative trait refers to the phenotypic characteristic that varies in degree and can be attributed to the interactions between two or more genes and their environment. There are three types of quantitative traits: continuous traits which have a continuum of possible phenotypes, meristic traits and discrete (or threshold) traits. Quantitative Trait Loci (QTL) are genes (or DNA regions that contain those genes) that affect quantitative trait variation in a population. QTL mapping is the statistical inference of the number and the genomic positions of QTL, and the relationship between QTL and phenotypic values of the corresponding quantitative trait.

Since 1908s, there has been a great deal of interest in the development of methodology to map QTL using the data from experimental crosses. A traditional experimental cross starts with two parental inbred lines, P_1 and P_2, with different trait values and different genotypes in genetic markers. Genetic markers contain information about segregation of a genome at various positions in a population. Examples of genetic markers include restriction fragment length polymorphism (RFLP), randomly amplified polymorphic DNA

(RAPD), amplified fragment length polymorphism (AFLP), microsatellite and single nucleotide polymorphism (SNP). Suppose the genotype of a genetic marker is AA in P_1 and aa in P_2, where A and a are two alleles for that genetic marker. Then their progeny F_1, heterozygous for all marker alleles, has genotype Aa for that marker. Typically, there are two types of experimental crosses, backcross ($F_1 \otimes P2$) and F_2 intercross ($F_1 \otimes F_1$). The segregating progeny have genotype Aa or aa for backcross, and AA, Aa or aa for F_2 intercross.

Experimental crosses are used to test for associations between genetic markers and the phenotypic trait of interest, to locate QTL that influence the quantitative trait. Let y_i be the measure of the phenotypic trait (assume continuous trait for simplicity) and x_{ij} be the genotype at marker j for individual i ($i = 1, 2, ..., n$). x_{ij} is coded as an indicator variable equal to the number of A alleles, which is 0 or 1 for backcross and 0, 1 or 2 for F_2 intercross. The relationship between the measure of phenotype y_i and the genotype x_{ij} can be modeled by marker regression, for the backcross case, as

$$y_i = a + bx_{ij} + \epsilon_i,$$

where ϵ_i is the random error independent identically distributed as $N(0, \sigma^2)$, and a, b and σ^2 are unknown parameters. Here, b denotes the effect of a single allele substitution on the phenotype. For the F_2 design, the model becomes

$$y_i = a + bx_{ij} + dz_{ij} + \epsilon_i,$$

where z_{ij} is a (0, 1)-indicator variable for dominance (1 for homozygote and 0 for heterozygote), and the parameter d is the dominant effect. The likelihood function, for the backcross design as an example, can be expressed as

$$L(a, b, \sigma^2) = \mathbf{\Pi}_i \phi(y_i - (a + bx_{ij}), \sigma^2),$$

where $\phi(y, \sigma^2) = (2\pi\sigma^2)^{-1/2} exp(-y^2/2\sigma^2)$ is the probability density function for a normal random variable with mean 0 and variance σ^2. This model, sometimes called marker regression, is a regular ANOVA model, and the result of hypothesis test indicates the degree of linkage between the jth genetic marker and QTL. It can be repeated for all genotyped markers, with some control for multiple testing such as Bonferroni correction or permutation procedures. The method of marker regression is simple, easy to incorporate any covariances (such as polygenic effect or environmental effect), and can be easily extended to multiple regression to account for multiple loci. Also, it does not require any genetic map of markers. However, the method has some disadvantages. It makes use of imperfect information about QTL location, excludes individuals with missing genotype data, and has low power if the linkage between any marker and QTL is weak (for example, for sparse marker data).

Lander and Botstein (1989) first introduced the interval mapping technique and employed maximum likelihood method for analysis. They extended the method above to model a putative QTL at any given location and its phenotypic effect through a pair of flanking markers. For a putative QTL falling between markers, although its genotype x_i is not directly observed, its conditional distribution can be estimated from the two flanking markers. The likelihood function can be expressed as

$$L(a, b, \sigma^2) = \mathbf{\Pi}_i \{p_i \phi(y_i - (a + b), \sigma^2) + (1 - p_i)\phi(y_i - a, \sigma^2)\},$$

where p_i is the probability that $x_i = 1$ conditional on the genetic location of the putative QTL and the genotypes of the flanking markers of the individual i. p_i is a function of the genotypes of the two flanking markers and the genetic distance between the putative QTL and the two flanking markers. The maximum likelihood estimator of the parameters can be obtained by EM algorithm (Lander and Botstein 1989).

The evidence of QTL can be indicated by the LOD score defined as

$$LOD = log_{10}\frac{L(\hat{a}, \hat{b}, \hat{\sigma^2})}{L(\tilde{a}, 0, \tilde{\sigma^2})},$$

where $(\hat{a}, \hat{b}, \hat{\sigma^2})$ is the unconstrained MLE, and $(\tilde{a}, 0, \tilde{\sigma^2})$ is the MLE restricted to $b = 0$. The LOD score is proportional to the likelihood ratio statistic (LR)

$$LOD = \frac{LR}{2log10} = 0.217LR,$$

where

$$LR = 2log\frac{L(\hat{a}, \hat{b}, \hat{\sigma^2})}{L(\tilde{a}, 0, \tilde{\sigma^2})}.$$

Under the null hypothesis, at each genetic location, the likelihood ratio statistic is asymptotically distributed as χ^2_1 for the backcross design and χ^2_2 for the F_2 intercross. Hence the asymptotic distribution for the LOD score is proportional to a chi-square random variable. This test can be performed at any position covered by markers and thus creates a systematic strategy of searching for QTL within the whole genome. The point estimate of the QTL position is the map position where the highest LOD was reached.

However, the strategy above introduces a multiple testing problem, and the distribution of the maximum LOD score over the whole genome is very complicated since the tests are correlated. For the sparse-map case where the markers are widely separated, one can assume marker intervals are approximately independent and apply Bonferroni correction to get the threshold of test statistic. For dense-map case, Lander and Botstein (1989) proposed to set the LOD threshold as $t_\alpha/(2log10)$, where t_α solves the equation

$$\alpha = (C + Gt_\alpha)\chi^2(t_\alpha),$$

where α is the significance level, C is the number of chromosomes, G is the genetic length

measured in Morgan, and $\chi^2(t_\alpha)$ denotes the cumulative distribution function of the chi-square distribution with 1 degree of freedom (for backcross). Lander and Botstein (1989) suggested a typical threshold of LOD score to be between 2 and 3, for approximately 5 percent overall false positive error in detecting QTL.

Churchill and Doerge (1994) proposed a permutation procedure, originated from Fisher's permutation test, to estimate empirical critical values for a given data set. This method starts by generating permuted samples of the data by randomly pairing marker genotypes with phenotypes. Then perform interval mapping analysis on those permuted samples to obtain an empirical distribution of the maximum LOD score. The $100(1-\alpha)$th percentile of the empirical distribution can be used as the threshold value.

1.1.2 QTL mapping for functional traits

Many quantitative traits, such as body size or weight, are inherently too complex to be described by a single value, because their phenotypes change with age, metabolic rate, environmental stimulus or other factors. These quantitative traits, which can be measured repeatedly over time, are called longitudinal traits, or infinite-dimensional characters by Kirkpatrick and Heckman (1989), or function-valued traits by Pletcher and Geyer (1999). For example, genetic correlations among age-specific weights in a laboratory population of rats were shown to involve variable gene action at different ages (Cheverud et al., 1983). Vaughn et al. (1999) located QTL responsible for age-specific weights in mice, and they found that some QTL affect the early growth patterns and some affect the late growth patterns.

A simple approach for mapping infinite-dimensional characters is to associate markers with phenotypes separately for different ages, traits, or environments and compare the difference of QTL expression across different categories, which is certainly inefficient because it fails to make full use of the information contained in the functional data.

Another approach is to treat the phenotypes measured from different time points as different traits and analyze the traits jointly using the method for multiple traits. However, as the number of traits increases, the multiple trait analysis approach will have a reduced ability to produce precise estimates of genetic parameters in quantitative genetic studies (Shaw 1987). Furthermore, if phenotypes are measured at different time points among subjects, it is impossible to apply the multiple trait analysis.

More recently, Wu and colleagues (Ma et al. 2002, Wu et al. 2002, Wu et al. 2004, Lin and Wu 2006) developed the functional mapping approach, which provided a useful framework for genetic mapping through mean and covariance modeling of longitudinal traits. They first used growth curve data as an example of functional traits, which is modeled by a parametric function such as sigmoidal or logistic function (Ma et al. 2002)

$$g(t) = \frac{a}{1 + be^{-rt}},$$

where t is the covariate such as time, $g(t)$ is the longitudinal trait at time t, a, b and r are parameters of the growth curve. The functional mapping method assumes that the set of parameters a, b and r is determined by the QTL genotype. Then the phenotypic trait of the ith individual of a backcross can be modeled as

$$y_i(t) = x_i g_1(t) + (1 - x_i) g_0(t) + \epsilon_i(t),$$

where x_i is a (0,1)-indicator variable for the QTL genotype of individual i, $g_j(t)$ is the growth curve if the QTL has genotype j ($j = 0$ or 1), and $\epsilon_i(t)$ is the random error. $\boldsymbol{\epsilon}_i = (\epsilon_i(t_1), \epsilon_i(t_2), ..., \epsilon_i(t_m))$ is assumed to be identical among different genotypes and follows a multivariate normal distribution, $N(0, \boldsymbol{\Sigma})$, with the covariance $\boldsymbol{\Sigma}$ modeled by

a first-order stationary autoregressive model (AR(1)),

$$\mathbf{\Sigma} = \sigma_e^2 \begin{bmatrix} 1 & \rho & \dots & \rho^{m-1} \\ \rho & 1 & \dots & \rho^{m-2} \\ \vdots & \vdots & \vdots & \vdots \\ \rho^{m-1} & \rho^{m-2} & \dots & 1 \end{bmatrix},$$

where m is the number of repeated measures of individual i. The maximum likelihood estimates of the unknown parameters can be computed by EM algorithm and likelihood ratio test can be employed for hypothesis testing. The Nelder-Mead simplex algorithm (Nelder and Mead 1965) can be used as an alternative to reduce the computational burden from the traditional EM algorithm (Zhao et al. 2004).

An important issue in longitudinal analysis is how to model the correlation among random errors of the same subject. There are several correlation structures that are widely used in longitudinal data modeling, such as independent, m-dependent, exchangeable and autoregressive. Those correlation structures are mostly simple with only a few parameters. For example, the stationary AR(1) model above involves only 2 parameters. However, the stationary AR(1) model assumes the longitudinal data has stationary variance and covariance, which is questionable in a lot of cases (Zhao et al. 2005).

To deal with the heteroscedastic problem of the residual variance, one approach is to model the residual variance by a parametric function of time (Pletcher and Geyer 1999). But this approach needs to implement additional parameters for characterizing the age-dependent change of the variance. Another approach is to use transform the data by the transform-both-sides method (Wu et al. 2004) and then use the AR(1) model on the transformed data to achieve stationary variance. However, the stationary covariance assumption is still a problem.

Antedependence models are important models in genetic studies (Jaffrezic et al. 2004). They are generalizations of stationary autoregressive models that are able to

model both nonstationary variance and correlation functions. The antedependence model was originally proposed by Gabriel (1962), which assumes serial correlation within subjects like the autoregressive model but allows for nonstationary variation. It states that an observation at a particular time t depends on the previous ones, with the degree of dependence decaying with time lag. If an observation at time t is independent of all observations before $t-r$, this antedependent model is called rth-order, or AD(r). A T-variate normal random vector $\mathbf{y} = (y_1, ..., y_T)^T$ with mean $\boldsymbol{\mu} = (\mu_1, ..., \mu_T)^T$ follows an AD(r) model if

$$\begin{array}{l} y_1 = \mu_1 + \epsilon_1 \\ y_t = \mu_t + \sum_{k=1}^{r^*} \phi_{kt}(y_{t-k} - \mu_{t-k}) + \epsilon_t \quad t = 2, 3, ..., T \end{array} ,$$

where $r^* = min(r, t-1)$, the ϵ_t's are independent random variables following $N(0, \sigma_t^2)$, and ϕ_{kt}'s are such that the covariance matrix, $\Sigma = \{\sigma_{ij}\}$, is positive definite. It is easy to observe that both variance and correlation are nonconstant over time, as long as some $\phi_{kt} \neq 0$. Antedependence models are very useful for longitudinal data exhibiting heterogeneous variances and nonstationary serial correlation, such as data in growth studies (Nunez-Anton and Zimmerman, 2000). However, the covariance of an unstructured AD(r) model, UAD(r), is specified by $(r+1)(2T-r)/2$ parameters, which is not so parsimonious.

To make the antedependence model more parsimonious and useful, Nunez-Anton (1997) and Nunez-Anton and Zimmerman (2000) proposed structured antedependence (SAD) models, which incorporate some structural forms of nonstationary into AD models. Denoting the measurement times of any subject as $t_1 < t_2 < ... < t_T$, an rth-order SAD, or SAD(r), model can be specified as

$$\begin{array}{l} \phi_{i-k,i} = f(t_i, t_{i-k}; \lambda_k) \\ \sigma_{ii} = \sigma^2 g(t_i; \psi) \end{array} ,$$

where $\sigma^2 > 0$, ψ, $\lambda_1, ..., \lambda_r$ are parameters such that the covariance matrix Σ is positive definite, $f(.)$ and $g(.)$ are specified functions. Nunez-Anton and Zimmerman (2000) suggested a typical choice of $f(.)$ can be an exponential function as

$$\phi_{i-k,i} = f(t_i, t_{i-k}; \lambda_k) = exp\{-\lambda_k(t_i - t_{i-k})\}.$$

$g(t_i; \psi)$ is a function of relatively few parameters (e.g. a low-order polynomial). Instead of using the function $g(t_i; \psi)$ to model the innovation variance changing with time, Nunez-Anton and Zimmerman (2000) suggested to model the $log\sigma^2(t_i)$ by some polynomial functions of t_i, as

$$\sigma^2(t_i) = exp\{a + bt_i\} = \sigma^2 exp\{bt_i\},$$

where $\sigma^2 = e^a$, a and b are unknown parameters. With such choices of $f(.)$ and $g(.)$, only $r + 2$ parameters are involved in the model, regardless of times of measurement T. Therefore, the SAD models are much more parsimonious than the UAD models when T is not too small.

Another good property of the SAD model is that it is easy to get the maximum likelihood estimators for the parameters for the model, because the inverse of the covariance matrix is easy to compute. Use SAD(1) model as an example, the residual covariance matrix $\mathbf{\Sigma}$ can be expressed as

$$\mathbf{\Sigma} = \mathbf{AGA}^T,$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \phi_{1,2} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \phi_{1,T} & \phi_{2,T} & \dots & 1 \end{bmatrix},$$

and

$$\mathbf{G} = \begin{bmatrix} \sigma^2(t_1) & 0 & \dots & 0 \\ 0 & \sigma^2(t_2) & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma^2(t_T) \end{bmatrix}.$$

The inverse of the matrix $\mathbf{A}$ is

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ -\phi_{1,2} & 1 & 0 & \dots & 0 & 0 \\ 0 & -\phi_{2,3} & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & -\phi_{T-1,T} & 1 \end{bmatrix}.$$

Hence it is easy to compute the inverse of $\mathbf{\Sigma}$.

The SAD(1) model can be further simplified, by assuming times of repeated measurements are equally spaced and innovation variances are constant over time points as introduced by Jaffrezic et al. (2003), and the measurements can be expressed as

$$\begin{aligned} y_1 &= \mu_1 + \epsilon_1 \\ y_t &= \mu_t + \phi(y_{t-1} - \mu_{t-1}) + \epsilon_t \quad t = 2, 3, ..., T \end{aligned},$$

where ϵ_t follows $N(0, \sigma^2)$ with constant innovation variance σ^2. This simple SAD(1) model only involves two parameters σ^2 and ϕ, and hence it is very parsimonious. The analytical forms for variance and covariance functions of this model can be derived as

$$\begin{aligned} \sigma_{ii} &= \tfrac{1-\phi^{2i}}{1-\phi^2}\sigma^2 \\ \sigma_{i-k,i} &= \phi^k \tfrac{1-\phi^{2(i-k)}}{1-\phi^2}\sigma^2 \end{aligned}.$$

It can be easily seen that both variance and correlation functions are non-stationary for

the SAD(1) model, even with constant innovation variance σ^2 and constant antedependent coefficient ϕ.

In many situations, we have little information about the correlation structure of random error. The generalized estimation equations (GEE) approach (Liang and Zeger 1986) provides a unified way to fit regression models with longitudinal data. Let y_{ij} be the response of the jth observation of the ith individual, with $E(y_{ij}) = \mu_{ij} = g(\mathbf{x}_{ij}^T\boldsymbol{\beta})$ for a link function $g()$. Let $\mathbf{A}_i$ be a diagonal matrix with elements $Var(\mathbf{y}_i)$ and $\mathbf{D}_i$ be a matrix with $\partial\mu_{ij}/\partial\beta_k$ being the element in the ith row and the jth column. The GEE estimates $\boldsymbol{\beta}$ by solving the following set of generalized estimation equations

$$\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{A}_i^{-1/2} \mathbf{R}_i^{-1} A_i^{-1/2} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0,$$

where $\mathbf{R}_i$ is the working correlation matrix for $\mathbf{y}_i = (y_{i1}, ..., y_{in})^T$ (n is the number of subjects) and $\boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{in})^T$. $\mathbf{R}_i$ is called working correlation matrix because it is not required to be correctly specified for the parameter estimates and their variance to be consistent, although there can be important gains in efficiency if correctly specifying the working correlation matrix (Liang and Zeger 1986). The GEE estimator is robust to the choice of working correlation as long as the number of subjects n is not too small, therefore the GEE method is extremely useful when the actual correlation structure is unknown.

1.1.3 Varying coefficient models

In many situations, the dynamic pattern of genetic effects has no obvious functional form. So it is desirable to have a more flexible way for modeling such genetic effects. Varying coefficient models were introduced by Cleveland et al. (1991), and discussed by Hastie and Tibshirani (1993) in more details, to extend the applications of local regression techniques from one-dimensional to multidimensional setting. A varying coefficient

model has the form

$$\mathbf{y} = \sum_{j=1}^{p} \beta_j(\mathbf{t})\mathbf{x}_j + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ is the response, $\boldsymbol{\epsilon}$ is the random error, and the covariate $\mathbf{t}$ changes the coefficients of the covariates $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p$ through the functions $\beta_1, \beta_2, ..., \beta_p$. Varying coefficient models are natural extensions of classic parametric models by relaxing the linear assumptions imposed on traditional parametric models and exploring the hidden structure. With good interpretability, they are becoming more and more popular. There are many ways to model the unspecified functions $\beta_j(\mathbf{t})$, such as polynomials, Fourier series, piecewise polynomials or more general nonparametric functions (Hastie and Tibshirani 1993).

Among those, the most straightforward one is the polynomials. One common choice of polynomials is Legendre polynomials, the simple classical orthogonal polynomials. Legendre polynomials have been extensively used by animal geneticists and breeders to fit milk production and other dynamic traits (Kirkpatrick and Heckman 1989; Kirkpatrick et al. 1990; Schaeffer 2004). By choosing different orders of orthogonal polynomials, the Legendre function has potential to approximate the functional relationships between trait values and times to any specified degree of precision. The general form of a Legendre polynomial of order r, defined over the interval [-1, 1], is given by the sum,

$$P_r(t) = \sum_{k=0}^{K} (-1)^k \frac{(2r-2k)!}{2^r k!(r-k)!(r-2k)!} t^{r-2k},$$

where $P_r(t)$ denotes the Legendre polynomial of order r and $K = r/2$ or $(r-1)/2$ whichever is an integer. It is easy to verify the Legendre polynomials above satisfies the property of orthogonal polynomials that they are all orthogonal over [-1, 1]; whenever $m \neq n$,

$$\int_{-1}^{1} P_n(t)P_n(t)dt = 0.$$

In this modeling, t is rescaled to [-1, 1] from the original measurement t^0 by

$$t = \frac{2(t^0 - t_{min})}{t_{max} - t_{min}} - 1,$$

where t_{min} and t_{max} are the first and last time points, respectively.

A more flexible way to model varying coefficients is to use nonparametric functions. For example, Hastie and Tibshirani (1993) introduced the local linear regression method to model $\beta_j(t)$. For a given t_0 and t close to t_0, $\beta_j(t)$ can be approximated by Taylor expansion

$$\beta_j(t) \approx \beta_j(t_0) + \beta_j'(t_0)(t - t_0) = b_j + c_j(t - t_0).$$

The parameter $\{b_j, c_j\}$ can be solved by solving the least-squares objective function

$$\sum_{i=1}^{n} \{y_i - \sum_{j=1}^{p} b_j + c_j(t - t_0)x_{ij}\}^2 K_h(t - t_0),$$

where $K_h(.)$ is a kernel function. It can be easily generalized to the local polynomial regression (Fan and Gijbels 1996). The local regression method needs to solve many weighted regression problems as the choice of t_0 is usually in the order of 100 (Fan et al. 2000). It assumes that the coefficient functions $\beta_j(t)$'s have similar degree of smoothness so that it can be equally well approximated in a local neighborhood. When the functions have different degrees of smoothness, Fan and Zhang (2000) showed that the local regression estimator is suboptimal under their asymptotic formulation.

Another nonparametric approach is the smoothing splines (Hastie and Tibshirani 1993; Hoover et al. 1998). It minimizes the penalized least-squares criterion

$$\sum_{i=1}^{n} \{y_i - \sum_{j=1}^{p} \beta_j(t)x_{ij}\}^2 + \sum_{j=1}^{p} \lambda_j \int \beta_j''(t)^2 dt,$$

where λ_j's are positive regularization parameters. This method is computationally intensive when there is a large number of distinct time points (Huang et al. 2002). Also,

it may suffer from the same problem as the local regression estimator when $\beta_j(t)$'s have different degrees of smoothness (Fan et al. 2000).

Huang et al. (2002) proposed to model the varying coefficient functions by function approximation through basis expansion, in which the coefficient $\beta_j(t)$ can be approximated by

$$\beta_j(t) \approx \sum_{k=1}^{K} \gamma_{jk} B_{jk}(t),$$

where $B_{jk}(t)$'s are basis functions and γ_{jk}'s are corresponding coefficients. Various basis systems can be used for the approximation, and the most common choice is the B-spline basis (He and Shi 1998, Huang et al. 2004, Wang et al. 2007, Wang et al. 2008). The smoothness of the coefficient functions modeled by B-splines are controlled by the number $K = n_j + d + 1$, where n_j is the number of interior knots and d is the degree of spline. The interior knots of the splines can be either equally spaced or placed on the sample quantiles of the data, so that there are about the same number of observations between any two adjacent knots. If equally spaced knots are used, the bases $B_{jk}(t)$'s are predetermined for any given t. Yang et. al (2009) applied B-splines in nonparametric functional mapping of QTL and estimating the underlying functional form of phenotypic trajectories. And they found the nonparametric method performs better than parametric methods.

1.1.4 QTL mapping by recombinant inbred intercrosses

QTL mapping in humans has a lot of difficulties: time consuming, expensive, hampered by ethical problems, compromised by small sizes, genetically diverse and subject to uncontrollable environments. All those obstacles can be overcome in the laboratory mice. Furthermore, most human genes have functional mouse counterparts and both genomes are organized similarly. Hence, the laboratory mouse has become a very important model organism in QTL mapping. However, the most widely used experimental

mapping approaches, particularly intercrosses and backcrosses, lack the genetic reproducibility to efficiently perform multi-variant analyses across traits and environmental conditions (Darvasi 1998). This is a particularly acute problem when one wants to examine numerous gene-environment interactions or study disease progression at many stages and ages (Zou et al. 2005). Experimental crosses can be replicated with recombinant inbred (RI) lines. Recombinant inbred lines are important resources that have contributed to genetic dissection of simple and complex traits. A major advantage of RI panels over other commonly used mapping approaches is their ability to support genetic mapping and correlations among many traits, even under different environmental conditions (Plonmin et al. 1991). However, mouse RI panels generally have low power and precision compared to other resources because of their small size; typical mouse RI panels have only 15-35 strains from a single pair of parental inbred lines (Zou et al. 2005).

A novel derivative of RI lines, called recombinant inbred intercrosses (RIX) has recently been designed, that permits repeated interrogation of a fixed, but complex genotype to reduce non-genetic variance while increasing the power of the original RI panel (Threadgill et al. 2002). The recombinant inbred intercrosses (RIX) panel is created as F1 intercrosses of the Collaborative Cross (CC) recombinant inbred lines. The Collaborative Cross is a large panel of new inbred mouse strains that are derived from an eight way cross using a set of founder strains including three wild-derived strains. A special breeding approach is designed to randomize the genetic makeup of each inbred line to create a panel of CC RI lines. Since all CC RI mice are homozygote at each locus, the genotypes of the derivative RIX mice will be known in advance by imputing from the genotypes of the parental CC lines. RIX mice with identical genotypes can be re-generated whenever needed. Hence, it can be used to study genetic effects of QTL under different environments or stages.

Compared to RI, the RIX has several advantages that includes twice the number of

recombination sites in a single individual since each is derived from two parental RI, albeit there are no new recombination sites; dominance effects can be estimated; a large expansion of different RIX genomes over the parental RI; and, because of the buffering capacity of their heterogeneous genome structure, RIX genomes should provide more reliable trait means than the parental RIs. The RIX approach also has advantages over classical crosses like the F_2 design since each RIX has a higher recombination density because of the map expansion of the parental RI, averaging almost four-fold more recombination sites than a single F_2 individual when performing interval mapping; the genotypes will be known in advance by imputing from the parental RI lines; RIX are especially useful for long-term collaborative research because their genotypes are renewable making the phenotypic data cumulative within the research community; and since RIX genomes are easily replicated, experiments with different environmental variables or temporal relationships can be performed on the same genotypes (Zou et al. 2005).

At individual level, although the genome of each RIX mouse has similar genetic structures of F_2 individuals, statistical analyses for F_2 data cannot be directly applied to RIX data. Because some RIX individuals share a common parental RI line, making them genetically more related to each other than those that do not share any parental lines. We need to handle this special correlation among subjects in data analysis. For example, for a set of data with L CC RI lines, there are at most $L(L-1)/2$ nonreciprocal RIXs that can be generated, which is a huge number when L is large. A useful sampling and mating scheme is the loop design as described by Zou et al. (2005). It starts by randomly ordering L RI lines to form a circle. Then each RI line is mated with the next J RI lines after it. That is, we mate RI_1 with RI_2, RI_3, ... and RI_{J+1}; ...; RI_i with $RI_{m(i+1,L)}$, $RI_{m(i+2,L)}$, ... and $RI_{m(i+J,L)}$; ...; and RI_L with RI_1, RI_2, ... and RI_J, where

$$m(x, L) = \begin{cases} x, & \text{if } x \leq L \ ; \\ x - L, & \text{if } x > L \ . \end{cases}$$

Assume that the trait of interested is affected by one major QTL and polygenes, a mixed effect model (Zou et al. 2005) can be employed where the QTL effect is treated as a fixed effect and the polygenic effect is treated as a random effect. The mixed model has the form

$$y_i = \mu + \beta_j x_{ij} + \sum_{l=0}^{L} a_{il}\alpha_l + \epsilon_i,$$

where y_i is the measure of the genotype of individual i at time t_i; μ is the overall population mean; x_{ij} is the genotype of the ith individual at the jth putative QTL, coded as -1, 0 or 1 for genotypes aa, Aa and AA, respectively; β_j is the effect of the jth putative QTL; the random polygenic effect α_l follows $N(0, \sigma_a^2)$ for $l = 1, 2, ..., L$; the random error ϵ_i follows $N(0, \sigma_0^2)$; and

$$a_{il} = \begin{cases} 1, & \text{if one of } ith \text{ individual's parents is } RI_l; \\ 0, & \text{otherwise .} \end{cases}$$

1.2 Multiple mapping and model selection

1.2.1 Multiple QTL mapping

The QTL mapping methods discussed above all assume that the quantitative trait of interest is affected by one major QTL. However, a large amount of traits in nature are affected by many genes (Zeng 1994). The interval mapping method works well if there is only one segregating QTL on a chromosome. However, when there is more than one QTL on a chromosome, the test statistic at one position will be affected by all those QTL. As a result, the estimates are likely to be biased and the mapping power may be decreased (Knott and Haley 1992).

The interval mapping method was improved by Zeng (1993, 1994) and Jansen and Stam (1994) in their method of composite interval mapping (CIM) by treating other intervals as covariates to control the overall genetic background. Consider a QTL between

markers j and $j+1$ the model can be expressed as

$$y_i = a + bx_i^* + \sum_{k \neq j, j+1} b_k x_{ik} + \epsilon_i,$$

where x_i^* is an indicator variable (taking value 0 or 1) for the genotype of the putative QTL located between makers j and $j+1$, b is the effect of the putative QTL, x_{ik} is the (0,1)-coded genotype of the kth marker of the ith individual, and b_k is the partial regression coefficient for the kth marker. The likelihood function can be expressed in the same form as that in the interval mapping, except replacing bx_i by $bx_i^* + \sum_{k \neq j, j+1} b_k x_{ik}$. Similarly, the maximum likelihood estimator of the parameters can be calculated using the expected conditional maximization (ECM) algorithm (Zeng 1994).

The composite interval mapping method is able to search for the number, positions and effects of QTL. However, it is nontrivial to determine the genome-wide significance for CIM. Furthermore, it does not consider the interaction between QTL. Kao and Zeng (1997) proposed multiple interval mapping (MIM), which makes use of the model

$$y_i = a + \sum_{j=1}^{m} b_j x_{ij}^* + \sum_{j \neq k}^{m} \delta_{jk}(w_{jk} x_{ij}^* x_{ik}^*) + \epsilon_i.$$

Here a is the overall mean. b_j is the marginal effect of QTL j. x_{ij}^* is an indicator denoting the genotype of putative QTL j for subject i, which is unobserved but can be inferred from marker data in terms of probability. δ_{jk} is an indicator for epistasis between QTL j and QTL k, whose value is 0 if there is no epistasis between the two QTL. w_{jk} is the epistatic effect between QTL j and QTL k. The likelihood function for this model is given by

$$L(\mu, \mathbf{a}, \boldsymbol{\delta}|) = \mathbf{\Pi}_{i=1}^{n} [\sum_{j=1}^{2^m} p_{ij} \phi(y_i - \mu_{ij}, \sigma^2)],$$

where p_{ij} is the joint conditional probability of 2^m possible genotypes of m putative QTL, and μ_{ij} is the expected phenotype for the combined genotypes. Similarly to the

CIM model, the maximum likelihood estimates can be obtained by ECM algorithm.

There are several advantages for the multiple interval mapping. The multiple interval mapping method is able to simultaneously search for number, positions, effects and epistatic interaction of significant QTL by selecting the best genetic model. It improves statistical power to identify QTL and improves the precision of estimating QTL position. Furthermore, it helps to understand the architecture of quantitative traits (Kao et al. 1999). However, the evaluation of MIM model is computationally intensive due to the high dimension of unknown parameters, especially when it is performed on the whole genome (Zeng et al. 2000). Hence, Kao et al. (1999) suggested a stepwise procedure. It selects a premodel before doing MIM analysis. Model selection methods for multiple regression, such as backward stepwise regression, can be applied to select a subset of markers and epistatic terms. After a premodel with reasonable size is picked, MIM analysis can be performed stepwisely to select the final best model, where the decision of dropping or retaining a marker or epistasis effect depends on the results of the likelihood ratio test based on the MIM model.

Alternatively, identifying QTL responsible for variation in experimental crosses can be viewed as one problem of model selection (Broman and Speed 2002, Manichaikul et al. 2009). Broman and Speed (2002) proposed to identify QTL through stepwise regression and MCMC sampling with a modified BIC criterion, and concluded that it identifies more QTL than CIM under some conditions. Manichaikul et al. (2009) extended the model selection method to allow for selection of pairwise interactions among QTL.

1.2.2 Traditional methods for model selection

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ is a vector of n responses, $\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, ..., \mathbf{x}_n^T)^T$ is the design matrix of order $n \times p$ with $\mathbf{x}_i^T = (x_{i1}, x_{i2}, ..., x_{ip})$ being a vector of p predictors for the i^{th} observation, $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients possibly with some zero elements, and $\boldsymbol{\epsilon}$ is a $n \times 1$ random vector following n-dimensional multivariate normal distribution $N(\mathbf{0}, \sigma^2\mathbf{I})$. For variable selection, the predictors are often standardized so that $\sum_{i=1}^n x_{ij} = 0$ and $\sum_{i=1}^n x_{ij}^2 = 1$, $j = 1, 2, ..., p$.

A most common way to produce a predictive model is the ordinary least squares (OLS) method, which minimizes the residual sum of squares $\sum_{i=1}^n (y_i - \sum_{j=1}^p \beta_j x_{ij})^2$. Although the OLS fitting produces the best linear unbiased estimator, it does not always provide a satisfactory model in terms of prediction accuracy and interpretability. The variance of the predicted values can be very high due to over fitting and the interpretability is usually unconvincing since the OLS retains all predictors in the model without further actions. With too many variables in the model, it is difficult to understand which variables are really important to the response. Furthermore, for high-dimensional data where $p > n$, it is impossible to get an OLS solution, since the linear equation system has no unique solution. Therefore, it is essential to come up with some model selection methods that are able to produce more sparse models.

The model selection problem aims at selecting variables that are really important to the response, and estimating the coefficients corresponding to those variables. Traditional model selection methods, such as best subset regression and stepwise regression, keeps a subset of candidate predictive variables, eliminates the rest, and uses OLS to estimate the coefficients corresponding to the retained predictors. Candidate models, obtained from best subset regression or stepwise regression, can be evaluated by some measure of prediction accuracy, information criteria, or some other criteria.

Prediction accuracy can be measured by the expected prediction error (PE) or the mean squared prediction error (MSPE). For a regression fit $\mathbf{X}\hat{\boldsymbol{\beta}}$ at a new observation

$\mathbf{x}_0^T$, the expected PE is

$$\begin{aligned} E[PE(\mathbf{x}_0)] &= E[(\mathbf{x}_0^T\boldsymbol{\beta} + \boldsymbol{\epsilon}_0 - \mathbf{x}_0^T\hat{\boldsymbol{\beta}})^2] \\ &= \sigma^2 + [E(\mathbf{x}_0^T\hat{\boldsymbol{\beta}}) - \mathbf{x}_0^T\boldsymbol{\beta}]^2 + E[\mathbf{x}_0^T\hat{\boldsymbol{\beta}} - E(\mathbf{x}_0^T\hat{\boldsymbol{\beta}})]^2 \\ &= \sigma^2 + bias^2 + Variance. \end{aligned}$$

Here the first term is the irreducible error that cannot be avoided even if $\boldsymbol{\beta}$ is known, and it is not affected by whichever model was chosen. The second term is the squared bias, which is the amount by which the average of the estimate differs from $\mathbf{x}_0^T\boldsymbol{\beta}$, and the third term, the variance, is the expected squared deviation of $\mathbf{x}_0^T\hat{\boldsymbol{\beta}}$ from its mean. The combination of the last two components, called the mean squared prediction error or the model error (ME), are sometimes used, instead of the expected PE, as a measure of prediction accuracy.

However, PE and MSPE cannot be obtained directly in practice since the true $\boldsymbol{\beta}$ is never known. Mallows (1973) proposed a statistic C_p as a prediction accuracy criterion. The C_p of a candidate model with q ($q \leq p < n$) predictors is

$$C_p = \frac{RSS_q}{\hat{\sigma}_F^2} - n + 2q,$$

where $RSS_q = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{q}\hat{\beta}_j x_{ij})^2$ is the residual sum of squares from the OLS fit of the candidate model, and $\hat{\sigma}_F^2 = RSS_F/(n-p)$, with RSS_F being the residual sum of squares of the full model, is an estimate of σ^2. $\hat{\sigma}_F^2 C_p/n$ is an unbiased estimator of MSPE, hence to minimize C_p is approximately to minimize MSPE (Mallows, 1973). Therefore, the candidate model with the least C_p is the most desirable.

Another way to estimate the prediction accuracy of a model is through sample reuse procedures, such as cross-validation (CV) proposed by Picard and Cook (1984). The simplest kind of cross-validation method divides data into two parts so that one can fit any candidate model by OLS using one subset of data, called training set, and estimate

the prediction error using the fitted model and another subset of data, called validation set. The result of simple cross-validation depends heavily on how the data is divided. K-fold cross-validation is a good improvement to overcome this deficiency, where the dataset is divided into K parts and simple CV is performed K times. Each time, one of the K subsets is used as the test set and the other $K-1$ subsets are put together to form a training set. The K results then can be averaged to produce a single estimation of the prediction error, which substantially reduces its variance. An extreme case of K-fold CV is the leave-one-out cross-validation, which uses a single observation from the original sample as the validation data, and the remaining observations as the training data. The leave-one-out cross-validation is equivalent to a K-fold CV, in which K equals to the number of observations.

The cross-validation method, especially k-fold CV and leave-one-out CV, can be very intensive in computation. To reduce the computational burden, the generalized cross-validation (GCV) method first introduced by Craven and Wahba (1979) as an alternative to CV. The GCV criterion can be expressed as

$$GCV = \frac{RSS_q}{n(1 - q^*/n)^2},$$

where q^* is the effective number of parameters. Besides the advantage in computation, the GCV criterion has been found to possess several favorable properties (Golub et al. 1979; Li 1987).

Information criteria are based on likelihood or information measures. The most popular criteria are Akaike information criterion (AIC) and Schwarzs bayesian information criterion (BIC). AIC, first proposed Akaike (1970, 1974), is defined as

$$AIC = -2l + 2q,$$

where l is the log-likelihood of the candidate model with q predictors. Similar to AIC,

BIC (Schwarz 1978) replaces the penalty term $2q$ by $qlog(n)$ and has the form

$$BIC = -2l + qlog(n),$$

with n being the sample size. Model selection based on information criteria picks the candidate model with minimum AIC or BIC. Comparing to AIC, BIC prefers smaller models since usually $log(n) > 2$. Bayesian model selection is another category of model selection methods, which picks the model with the highest posterior probability among all candidate models. Beside the posterior probability, other criteria, such as Bayes factor or deviance information criterion (DIC), are frequently used in Bayesian model selection.

Based on any of those criteria, the best subset regression method exhaustively searches all possible subsets and selects the one whose criterion is optimal. For a full model with p predictors, there will be 2^p candidate models, hence the commotional burden for best subset regression is very high when p is large, and it is generally considered impractical for $p > 30$. To overcome this problem, stepwise regression searches through a smaller number of subsets. There are three strategies to do stepwise regression, forward selection, backward elimination and the combined method. Starting with no variables in the model, the forward selection method adds one predictor at a time, based on certain criterion, until a preset stopping threshold is met or all predictors are in the model. The forward selection method can be applied even when $p > n$. However, it may perform badly in presence of severe multicollinearity. On the other hand, the backward elimination method starts from the full model, eliminates one predict variable at a time, based on certain criterion, until a preset stopping threshold is met or all predictors are eliminated. The backward elimination method only works for $p < n$. The combined method performs both forward selection and backward elimination, testing at each stage for variables to be included or excluded. The stepwise regression usually finishes in less than p steps, much simpler than the best subset regression, and it is tunable since the

selection of elimination threshold can be adjusted. However, the method also has serious drawbacks. It is a discrete process, in which a coefficient is either set to zero or is inflated. The inherent discreteness makes results by stepwise regression highly variable, and unstable with respect to small perturbations in the data.

1.2.3 Model selection based on penalized likelihood

To overcome those drawbacks, a family of new variable selection methods has been proposed that are based on penalized likelihood. The common idea of this family of methods is to add a penalty function to the negative log-likelihood, or equivalently the residual sum of squares, which helps to shrink small components of $\boldsymbol{\beta}$ to zero when the objective function is minimized. These methods are different from traditional model selection methods in that they delete insignificant variables by estimating their coefficients as zero, and hence they are able to perform variable selection and parameter estimation simultaneously.

The first penalized least squares method, ridge regression (Horel and Kennard, 1970), however, is not designed to do variable selection. It retains all predictors in the model but modifies the way how the regression coefficients are estimated, by defining the estimator $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}} = argmin\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2\} \qquad \text{subject to } \sum_{j=1}^{p}\beta_j^2 \leq t,$$

where t is the tuning parameter. The optimization problem above is equivalent to

$$\hat{\boldsymbol{\beta}} = argmin\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2\},$$

where the tuning parameter λ has a one-to-one relation with t. Comparing to OLS, the estimates of ridge regression coefficients are shrunk towards zero. The degree of shrinkage depends on the tuning parameter, which can be selected using cross-validation method that is performed through a set of possible tuning parameters and the one with smallest

prediction error will be picked. At the price of introducing bias, the ridge estimators have lower variance, comparing to OLS, and more stable to small perturbation in the data. Moreover, ridge regression can be used to high dimensional data where $p > n$. However, ridge estimates do not shrink to zero, thus all predictors are retained in the model, which is not helpful with the interpretability.

Therefore, it is desirable to have a method that is able to shrink the estimates of unimportant regression coefficients to zero, and thus automatically select a subset of predictors. The non-negative garrote, proposed by Brieman (1995), is such a penalized regression method. Brieman's method minimizes

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} c_j \hat{\beta}_j^0 x_{ij})^2 \qquad \text{subject to } c_j \geq 0, \sum_{j=1}^{p} c_j \leq t,$$

where $\hat{\beta}_j^0$ is the OLS estimate and it is shrunk by a non-negative factor c_j to get the non-negative garotte estimate $c_j\hat{\beta}_j^0$, which can be shrunk to zero when c_j becomes zero. Brieman showed that the non-negative garotte has consistently lower prediction error than subset selection and is competitive with ridge regression except when the true model has many small non-zero coefficients. A drawback of this method is that its solution depends heavily on both the sign and the magnitude of the OLS estimates. Therefore, it suffers in overfit or highly correlated settings where the OLS estimates perform poorly. Also, for high dimensional data, non-negative garotte is not able to selection more predictors than the number of observations.

Motivated from the idea of non-negative garotte, Tibshirani (1996) introduced the least absolute shrinkage and selection operator (LASSO), which estimates the coefficients by

$$\hat{\boldsymbol{\beta}} = argmin\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|\}.$$

LASSO is able to produce sparse solutions, and thus select a parsimonious model. Tibshirani (1996) also provide a computational algorithm to get LASSO estimates by solving

the constrained least squared problem. Fu (1998) developed a "shooting algorithm" for LASSO. Efron et al. (2004) proposed a new model selection algorithm, the least angel regression (LARS), and showed that it turns to LASSO by some simple modification. The LARS algorithm simplified the implementation of LASSO. LASSO can be applied to high dimensional data, but it cannot select more variables than the number of predictors. Like ridge regression, LASSO estimates are also biased by shrinking toward zero. Furthermore, LASSO may not perform well for the data where predictors have very high correlation.

Both LASSO and ridge regression can be viewed as special cases of bridge regression introduced by Frank and Friedman (1993). Bridge regression minimizes

$$\hat{\boldsymbol{\beta}} = argmin\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j|^\gamma\} \qquad \text{with } \gamma \geq 1,$$

where γ is the shrinkage parameter. It becomes LASSO when $\gamma = 1$ and ridge regression when $\gamma = 2$. Fu (1998) developed a general algorithm to solve for bridge estimates and he showed, via simulation, that bridge regression outperforms LASSO and ridge regression in terms of reducing prediction error when the full model contains many coefficients that are either zero or large in absolute value. However, the bridge regression does not produce a sparse solution when $\gamma > 1$ (Fan and Li 2001).

It will be very appealing if a procedure that performs model selection and parameter estimation at the same time can achieve the sparsity and unbiasedness for large effects. Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) method, which makes use of a non-convex penalty function

$$p_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j|, & \text{if } |\beta_j| \leq \lambda; \\ -\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}, & \text{if } \lambda < |\beta_j| \leq a\lambda; \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |\beta_j| > a\lambda. \end{cases}$$

Here, a and λ are two tuning parameters that can be chosen by minimizing the estimated prediction error. Simulation studies showed that $a = 3.7$ generally works quite well for a variety of settings. Fan and Li (2001) emphasized that the SCAD estimator possesses three desirable properties: unbiasedness, sparsity and continuity. Fan and Li (2001) showed that under some minor regularity conditions, the SCAD estimator possesses the oracle property, that is, asymptotically performing as well as if the true submodel is known, while the LASSO estimator does not have the oracle property.

The SCAD penalty function is non-convex, so regular numerical methods cannot be directly applied to solve the objective function. Fan and Li (2001) proposed a unified algorithm for the minimization of penalized likelihood via local quadratic approximations, where the penalty function can be approximated by

$$\begin{aligned} p_\lambda(|\beta_j|) &\approx p_\lambda(|\beta_{j0}|) + [p_\lambda(|\beta_j|)]'|_{\beta_j=\beta_{j0}}(\beta_j - \beta_{j0}) + \frac{1}{2}[p_\lambda(|\beta_j|)]''|_{\beta_j=\beta_{j0}}(\beta_j - \beta_{j0})^2 \\ &\approx \frac{p'_\lambda(|\beta_{j0}|)}{|\beta_{j0}|}\beta_{j0}(\beta_j - \beta_{j0}) + \frac{1}{2}\frac{p'_\lambda(|\beta_{j0}|)}{|\beta_{j0}|}(\beta_j - \beta_{j0})^2 \\ &= \frac{1}{2}\frac{p'_\lambda(|\beta_{j0}|)}{|\beta_{j0}|}\beta_j^2. \end{aligned} \tag{1.1}$$

Thus the minimization of the objective function becomes the minimization of a quadratic function.

All the penalized least squares methods mentioned above add different penalty functions to the objective function of OLS procedure, and obtain new objective functions have the form

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|),$$

where $p_\lambda()$ is the penalty function depending on the tuning parameter λ. Minimizing the objective function above with respect to β_j's gives the penalized least squares estimator of $\boldsymbol{\beta}$. For example, the L_1 penalty function $p_\lambda(|\beta_j|) = \lambda|\beta_j|$ corresponds to LASSO; the

L_2 penalty function $p_\lambda(|\beta_j|) = \lambda\beta_j^2$ corresponds to ridge regression and the L_0 penalty, or entropy penalty, $p_\lambda(|\beta_j|) = \lambda I(\beta_j \neq 0)$ corresponds to variable subset selection.

The idea of applying penalty to the OLS criterion can be extended to penalized likelihood so that it can be employed by many likelihood-base models. The penalized log-likelihood takes the form

$$-l(\boldsymbol{\beta}) + \sum_{j=1}^{J} p_\lambda(|\boldsymbol{\beta}_j|),$$

where $l(\boldsymbol{\beta})$ the log-likelihood for $\boldsymbol{\beta}$. Under the setup of linear regression, the penalized least squares estimator and the penalized likelihood estimator are exactly the same for type 1 penalty functions defined by Zou and Li (2008), such as the bridge penalty functions.

The penalized likelihood estimators have a Bayesian interpretation, where the penalty functions can be thought of as log-prior densities for the parameters. For example, the LASSO estimator can be derived as Bayes posterior mode under independent double-exponential priors for the regression coefficients β_j's (Tibshirani 1996), the ridge estimator is the posterior mode under independent Gaussian priors for parameters, and the SCAD estimator can be viewed as the posterior mode under improper priors.

The penalized least squares methods, like LASSO, are very appealing in variable selection. However, LASSO has some drawbacks. For high dimensional data, LASSO cannot select more variables than the number of predictors. Furthermore, in the setup of highly correlated predictors, LASSO tends to pick only one variable within a group of highly correlated predictor variables. In the case of high correlation between predictors, the prediction performance of LASSO has been found to be dominated by ridge regression (Tibshirani 1996). Zou and Hastie (2005) proposed a new variable selection method, the elastic net, to retain good features of both LASSO and ridge regression. The naive elastic

net criterion is

$$\hat{\boldsymbol{\beta}} = argmin\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j^2\},$$

equivalent to minimize

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 \qquad \text{subject to } (1-\alpha)\sum_{j=1}^{p}|\beta_j| + \alpha\sum_{j=1}^{p}\beta_j^2 \leq t,$$

where $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2} \in [0,1)$ with λ_1 and λ_2 being two tuning parameters. The elastic net penalty is the convex combination of the LASSO penalty and the ridge penalty. The naive elastic net overcomes the limitation of LASSO in both high dimensional data and the setting of highly correlated predictors. However, it generally does not perform well in prediction due to the double shrinkage. To correct this deficiency, the naive elastic net estimator is rescaled to get the elastic net estimator as

$$\hat{\boldsymbol{\beta}}\{elasticnet\} = (1+\lambda_2)\hat{\boldsymbol{\beta}}\{naiveelasticnet\}.$$

Zou and Hastie (2005) extended the LARS algorithm and developed the LARS-EN algorithm to solve the elastic net estimate. Simulation results showed that under collinearity, the elastic net dominates LASSO in terms of prediction and it exhibits "grouped section" ability by selecting more variables than LASSO.

Another drawback of LASSO is that the LASSO estimator is not consistent. Zou (2006) showed that if the value of a regression coefficient is zero, the probability that its LASSO estimate is zero is generally less than one. Hence, the LASSO is in general not variable selection consistent. He also provided a condition on the design matrix for the LASSO to be variable selection consistent. Zhao and Yu (2007) called this condition the irrepresentable condition on the design matrix. To improve the consistency of LASSO,

Zou (2006) proposed the adaptive LASSO, which has the form

$$\hat{\boldsymbol{\beta}} = argmin\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\hat{w}_j|\beta_j|\},$$

where $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\beta}}|^{\gamma}$, $\gamma > 0$, and $\hat{\boldsymbol{\beta}}$ is a root-n consistent estimator of $\boldsymbol{\beta}$. By incorporating data-dependent weights $\hat{\mathbf{w}}$, adaptive LASSO manages to reduce the bias of LASSO, especially when the true unknown parameter is large. Huang et al. (2006) showed if a reasonable initial estimator is available, then under appropriate conditions, the adaptive LASSO possesses the oracle property. In addition, under a partial orthogonality condition in which the covariates with zero coefficients are weakly correlated with the covariates with nonzero coefficients, the adaptive LASSO has the oracle property even if the number of covariates is much larger than the sample size, when using marginal regression to obtain the initial estimator.

Besides adaptive LASSO, a number of different variations of LASSO have been developed, to implement the LASSO procedure in different ways. For example, Tibshirani et al. (2005) introduced the fused LASSO to deal with those experiment designs where the predictors are ordered in some meaningful way, by penalizing the L_1-norm of both the coefficients and their successive differences,

$$\hat{\boldsymbol{\beta}} = argmin\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2\} \qquad \text{subject to } \sum_{j=1}^{p}|\beta_j| \le t_1 \text{ and } \sum_{j=2}^{p}|\beta_j - \beta_{j-1}| \le t_2.$$

The fused LASSO overcomes the drawback of ignoring the ordering of predictors, and it is especially useful when the number of predictive variables are much greater than the sample size.

To solve the problem of selecting grouped variables, for example, effects of factors in ANOVA, Yuan and Lin (2006) proposed the group LASSO method. The group LASSO

estimator is obtained by

$$\hat{\boldsymbol{\beta}} = argmin\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\boldsymbol{\beta}_j x_{ij})^2 + \lambda\sum_{j=1}^{J}(\boldsymbol{\beta}_j^T\mathbf{K}_j\boldsymbol{\beta}_j)^{1/2}\},$$

where $\boldsymbol{\beta}$ consist of J groups of regression coefficients, $\boldsymbol{\beta}_j$. $\mathbf{K}_j$s are some positive definite kernel matrices, for which a simple choice will be $\mathbf{K}_j$=$\mathbf{I}_{p_j}$, where $\mathbf{I}_{p_j}$ is the p_j-dimensional identity matrix with p_j being the number of regression coefficients in the jth group. Similarly, Yuan and Lin (2006) also extended LARS and non-negative garrote to group LARS and group non-negative garrote. To choose the tuning parameter λ, the authors introduced a C_p-type criterion, with approximated degrees of freedom

$$df = \sum_{j=1}^{J} I(||\boldsymbol{\beta}_j|| > 0) + \sum_{j=1}^{J}\frac{||\boldsymbol{\beta}_j||}{||\boldsymbol{\beta}_j^{LS}||}(p_j - 1),$$

where $\boldsymbol{\beta}_j^{LS}$ is the OLS estimate, for group LASSO. Empirical evidence suggested that the performance this C_p-type criterion is generally comparable with, and sometimes better than, that of fivefold cross-validation. Simulation results showed that these methods outperformed the traditional stepwise backward elimination method.

Group LASSO is designed for group selection, but it does not selection variables within any selected group. Huang et al. (2009) proposed the group bridge approach that is able to perform selection at the group and within-group individual variable levels simultaneously. The authors claimed that the group bridge possesses oracle group selection property, while the group LASSO does not.

Similar to group LASSO, Wang et al. (2007) developed the group smoothly clipped absolute deviation (SCAD) method, and applied it into a study to determine the transcriptional factors (TFs) involved in gene regulation during a biological process. They modeled time-varying effects of transcriptional factors with B-spline basis functions to form groups of predictors, and they concluded that the group SCAD regression is very effective for identifying variables with time-varying coefficients. Wang et al. (2008)

improved the method to selection both the group of variables and the number of basis functions within groups.

Penalized regression methods usually produce a set of candidate models with a grid of tuning parameters. An important issue for the penalized regression methods is how to select the best model from the set of candidate models, which is equivalent to how to select for the best tuning parameter. The most commonly used methods are cross-validation and generalized cross-validation. The CV method can be employed to estimate the prediction accuracy of a selected model based on a certain tuning parameter λ,

$$CV(\lambda) = ||\mathbf{y} - \mathbf{X}_0^T \hat{\boldsymbol{\beta}}||^2,$$

where $\hat{\boldsymbol{\beta}}$ is obtained from the training data set using the tuning parameter λ. $\mathbf{y}$ and $\mathbf{X}_0$ are from the validation data set, but the subset of variables in the design matrix $\mathbf{X}_0$ is determined using the penalized regression method on the training set. The tuning parameter λ can be selected from a predetermined set of values, by minimizing the $CV(\lambda)$. To overcome the deficient that the result depends heavily on how the data is divided, simple cross-validation can be easily extended K-fold cross-validation,

$$CV(\lambda) = \frac{1}{K} \sum_{k=1}^{K} ||\mathbf{y}_k - \mathbf{X}_k^T \hat{\boldsymbol{\beta}}_k||^2,$$

where $\mathbf{X}_k$ and $\mathbf{y}_k$ are from the kth subsample as validation set, and the columns of $\mathbf{X}_k$ and $\hat{\boldsymbol{\beta}}_k$ are obtained from the rest $K-1$ subsamples as training set. An extreme case of K-fold CV is the leave-one-out cross-validation, where K equals to the number of observations.

A drawback of k-fold CV and leave-one-out CV is that they can be very time consuming in computation. The generalized cross-validation method is an alternative to cross-validation that is faster in computation. Craven and Wahba (1979) generalized the

GCV criterion for ridge regression as

$$GCV(\lambda) = \frac{\frac{1}{n}||(\mathbf{I} - \mathbf{A}(\lambda)\mathbf{y})||^2}{[\frac{1}{n}Trace(\mathbf{I} - \mathbf{A}(\lambda))]^2},$$

where $\mathbf{A}(\lambda) = \mathbf{X}(\mathbf{X}^T\mathbf{X} + n\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$. Tibshirani (1996) extended the generalized cross-validation statistic to apply it to the LASSO procedure as

$$GCV(\lambda) = \frac{RSS_\lambda}{n(1 - df(\lambda)/n)^2},$$

where $df(\lambda) = trace\{\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{W}^-)^{-1}\mathbf{X}^T\}$ is the effective number of parameters, also known as the degrees of freedom, with $\mathbf{W} = diag(|\hat{\beta}_j|)$ for the LASSO estimate $\hat{\beta}$.

For linear models, GCV is asymptotically equivalent to C_p, AIC, and leave-one-out CV (Shao 1993, 1997; Hastie et al. 2001). However, GCV tends to overfit, hence Zou et al. (2007) proposed the BIC-LASSO shrinkage, which performs the LASSO model selection and chooses the tuning parameter λ by minimizing the BIC criterion

$$\frac{RSS_\lambda}{n\sigma^2} + \frac{log(n)}{n}df(\lambda).$$

The authors argued that the BIC criterion is more appropriate, comparing to AIC and C_p, when variable selection is the primary concern. Huang et al. (2009) made use of a slightly different form of BIC-type criterion

$$log(\frac{RSS_\lambda}{n}) + \frac{log(n)}{n}df(\lambda),$$

for group bridge method. They also compared several different tuning parameter selection methods and found that tuning based on BIC in general does better than that based on C_p, AIC or GCV in terms of selection at the group and individual variable levels.

Those criteria for choosing tuning parameter, including GCV, C_p, AIC and BIC, all involve estimating the effective degrees of freedom, $df(\lambda)$, which is an informative

measurement of model complexity. For simple linear models, the degrees of freedom is simply the number of predictors in the model. Unfortunately, due to the nonlinear nature of the LASSO, the explicit expression of the degrees of freedom, $df(\lambda)$, is not available. $df(\lambda)$ can be approximately estimated by $df(\lambda) = trace\{\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{W}^-)^{-1}\mathbf{X}^T\}$, as proposed by Tibshirani (1996). Besides that, Zou et al. (2007) stated that the easiest approach is to ignore shrinkage and use $df(\lambda) = q$, where q is the number of non-zero parameters. They also showed that it is an unbiased estimate of $df(\lambda)$ and this approximation is reasonable despite of its simplicity.

To perform model selection for longitudinal data, the penalized regression models can be extended to the penalized GEE models. Fu (2003) proposed a generalization of the bridge and LASSO penalties to GEE models. Similar to the penalized log-likelihood, for generalized linear models, the objective function for the penalized GEE models becomes the penalized deviance

$$Deviance + \sum_{j=1}^{J} p_\lambda(|\boldsymbol{\beta}_j|),$$

where $Deviance = 2l(\mathbf{y};\mathbf{y}) - 2l(\boldsymbol{\mu};\mathbf{y})$ (McCullagh and Nelder, 1989). To solve for the estimator of the penalty model, take the partial derivatives with respect to parameter β_j, leading to the following set of equations

$$\begin{cases} F_1(\boldsymbol{\beta}, \mathbf{X}, \mathbf{y}) + d(\beta_1) = 0 \\ ... \\ F_p(\boldsymbol{\beta}, \mathbf{X}, \mathbf{y}) + d(\beta_p) = 0, \end{cases}$$

where $F_j(\boldsymbol{\beta}, \mathbf{X}, \mathbf{y})$ is the jth score of the likelihood and $d(\beta_j)$ is the partial derivative of the penalty function respect to β_j. The set is equivalent to

$$\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) + d(\boldsymbol{\beta}) = 0,$$

where $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i\mathbf{A}_i^{1/2}$ and $d(\boldsymbol{\beta}) = (d(\beta_1), ..., d(\beta_p))^T$. Fu (2003) suggested solving the

above set by adjusting the iteratively reweighted least squares method for the penalty function. If assuming $\mathbf{R}_i$ is known, $\mathbf{V}_i$ can be computed and $\boldsymbol{\beta}$ can be solved just like regular penalized least squares. Then with solved $\boldsymbol{\beta}$, the parameters in $\mathbf{R}_i$ can be estimated. These steps can be performed iteratively to solve for both $\boldsymbol{\beta}$ and nuisance parameters. Fu (2003) showed that the penalized GEE potentially improves the performance of the GEE estimator and enjoys the same properties as linear penalty models.

The tuning parameter λ for penalized GEE models can be chosen by optimizing some criterion which balances goodness of fit and model complexity. Those criteria are adapted from classical criteria such as GCV, C_p, AIC and BIC. For example, Fu(2005) extended the GCV into the quasi-GCV (QGCV); Cantoni et al. (2005) proposed GC_p, a generalization of Mallows C_p (Mallows, 1973); Pan(2001) extended AIC into GEE setting, as the quasi-likelihood information criterion (QIC); Dziak and Li (2006) proposed two versions of generalized BIC criteria for penalized GEE models. Implementing such criteria is complicated for two reasons. First, because of combining shrinkage and selection, it is not clear how to define an effective degrees of freedom for the model. Second, because of clustered data, it is not clear how to measure goodness of fit or effective sample size.

1.3 Outlines for the thesis

The organization of the thesis is as follows. In Chapter 2, we introduce a nonparametric varying coefficient model to map QTL with phenotypes measured across multiple values of covariates for RIX data. We apply a nested permutation method to obtain threshold values. In Chapter 3, we extend the single QTL mapping to multiple QTL mapping via variable selection and propose a penalized mixed effects model. In Chapter 4, we expand the multiple QTL mapping method to longitudinal data. We propose a penalized likelihood method for longitudinal data with nonstationary covariance.

Chapter 2

Varying Coefficient Models for Mapping QTL Using RIX

2.1 Introduction

During the past two decades, there has been considerable development in statistical methodologies for mapping Quantitative Trait Loci (QTL), since Lander and Botstein (1989) implemented maximum likelihood approach to the interval mapping technique. The interval mapping method was later improved to composite interval mapping (CIM) by including markers from other intervals as covariates (Jansen and Stam 1994, Zeng 1994), and multiple interval mapping (MIM) by modeling all markers and their interactions (Kao and Zeng 1997). Besides the interval mapping approach, many other statistical approaches have been used in QTL mapping, such as regression analyses (Haley and Knott 1992) and the Bayesian approach (Yi and Xu 2000).

While these methods have been instrumental for QTL identification, they are not able to capture the temporal pattern of genetic effect. Many quantitative traits, such as body size, change with age, metabolic rate, environmental stimulus or other factors. These quantitative traits can be measured at different time points, ages or dosages, which allows us to study the pattern of genetic effects with the change of certain covariates. For

example, genetic correlations among age-specific weights in a laboratory population of rats were shown to involve variable gene action at different ages (Cheverud et al. 1983). Vaughn et al. (1999) located QTL responsible for age-specific weights in mice, and they found that some QTL affect the early growth patterns and some affect the late growth patterns. To study genetic determination of such functional traits, Wu and colleagues (Ma et al. 2002, Wu et al. 2002, Wu et al. 2004, Lin and Wu 2006) developed the functional mapping approach, which provided a useful framework for genetic mapping. They used growth curve data as an example of functional traits, and the genetic effect was modeled by a parametric function such as sigmoidal or logistic function (Ma et al. 2002). Alternatively, Zhang and Zhong (2006) proposed a variance components model for mapping functional traits, by modeling genetic effects as polynomial functions of time. While the parametric nature of functional mapping offers tremendous biological and statistical advantages, a reliance on the availability of mathematical functions limits its applicability (Yang et al. 2009).

The varying coefficient models are alternative statistical tools to explore dynamic patterns. The varying coefficient models were introduced by Cleveland et al. (1991), and discussed by Hastie and Tibshirani (1993) in more details, to extend the applications of local regression techniques from one-dimensional to multidimensional setting. In varying coefficient models, there are many ways to model the function of the varying effect, such as polynomials, Fourier series, piecewise polynomials or more general nonparametric functions (Hastie and Tibshirani, 1993). Hastie and Tibshirani (1993) introduced the local linear regression method to model varying coefficients. Fan and Zhang (2000) showed that the local regression estimator is suboptimal under their asymptotic formulation when the functions have different degrees of smoothness. Another nonparametric approach is the smoothing splines by minimizing a penalized least-squares criterion (Hastie and Tibshirani 1993; Hoover et al. 1998). However, the smoothing splines method is computationally intensive when there is a large number of distinct time points

(Huang et al. 2002), and it may suffer from the same problem as the local regression estimator when varying coefficient functions have different degrees of smoothness (Fan et al. 2000). Huang et al. (2002) proposed to model the varying coefficient functions by function approximation through basis expansion. Various basis systems can be used for the approximation, and the most common choice is the B-spline basis (He and Shi 1998, Pittman 2002, Huang et al. 2004, Wang et al. 2007, Wang et al. 2008). Comparing to some nonparametric approach like smoothing splines, one advantage of B-splines is that the smoothing matrix is independent of the responses. Yang et al. (2009) proposed a nonparametric functional mapping framework for genetic mapping of QTL controlling for a dynamic trait, implemented with B-splines.

The inbred mouse is a very important model organism in mapping QTL. QTL mapping in humans is difficult, time consuming, expensive, hampered by ethical problems, and compromised by populations that are too small, genetically diverse, and subject to uncontrollable environments. Those obstacles are nearly all overcome in the laboratory mouse. Furthermore, most human genes have functional mouse counterparts and both genomes are organized similarly. However, the traditional laboratory mice have a limited amount of variation (Darvasi 1998). This is a particularly acute problem when one wants to examine numerous gene-environment interactions or study disease progression at many stages and ages (Zou et al. 2005). Recombinant inbred (RI) lines are important resources that have contributed to genetic dissection of simple and complex traits. A major advantage of RI panels over other commonly used mapping approaches is their ability to support genetic mapping and correlations among many traits, even under different environmental conditions (Plonmin et al. 1991). However, mouse RI panels generally have low power and precision compared to other resources because of their small size; typical mouse RI panels have only 15-35 strains from a single pair of parental inbred lines (Zou et al. 2005).

The collaborative cross (CC) project (Threadgill et al. 2002) has been carried out

to create a large panel of new inbred mouse strains. It generates a large number of CC RI lines, from an eight way cross using eight founder strains, which makes the CC RI lines closer to nature population than regular RI lines by having more genetic variation. A novel derivative of RI lines, called recombinant inbred intercrosses (RIX) has been designed, that permits repeated interrogation of a fixed, but complex genotype to reduce non-genetic variance while increasing the power of the original RI panel (Threadgill et al. 2002). Since all CC RI mice are homozygote at each locus, the genotypes of the derivative RIX mice will be known in advance by imputing from the genotypes of the parental CC lines. RIX mice with identical genotypes can be re-generated whenever needed. At individual level, although the genome of each RIX mouse has similar genetic structures of F_2 individuals, statistical analyses for F_2 data can not be directly applied to RIX data. Because some RIX individuals share a common parental RI line, making them genetically more related to each other than those that do not share any parental lines. Compared to RI, the RIX has several advantages that includes twice the number of recombination sites in a single individual since each is derived from two parental RI; dominance effects can be estimated; a large expansion of different RIX genomes over the parental RI; and, because of the buffering capacity of their heterogeneous genome structure, RIX genomes should provide more reliable trait means than the parental RIs. The RIX approach also has advantages over classical crosses like the F_2 design since each RIX has a higher recombination density because of the map expansion of the parental RI, averaging almost four-fold more recombination sites than a single F_2 individual when performing interval mapping; the genotypes will be known in advance by imputing from the parental RI lines; RIX are especially useful for long-term collaborative research because their genotypes are renewable making the phenotypic data cumulative within the research community; and since RIX genomes are easily replicated, experiments with different environmental variables or temporal relationships can be performed on the same genotypes (Zou et al. 2005).

The remainder of the chapter is organized as follows. In the method section, we introduce a mixed effect model for RIX data with functional approximation for the genetic effect. Simulation studies are performed to evaluate the performance of varying coefficient model in the results section. Finally, in the discuss section, we summarize and discuss the implications of our model.

2.2 Methods

For a recombinant inbred panel with L lines, there are at most $L(L-1)/2$ nonreciprocal RIXs that can be generated (Zou et al. 2005), which is a huge number when L is large. A useful sampling and mating scheme is the loop design as described by Zou et al. (2005). With the loop design, L RI lines were randomly ordered to form a circle. Then each RI line is mated with the next J RI lines after it. That is, we mate RI_1 with RI_2, RI_3, ... and RI_{J+1}; ...; RI_i with $RI_{m(i+1,L)}$, $RI_{m(i+2,L)}$, ... and $RI_{m(i+J,L)}$; ...; and RI_L with RI_1, RI_2, ... and RI_J, where

$$m(x, L) = \begin{cases} x, & \text{if } x \leq L \ ; \\ x - L, & \text{if } x > L \ . \end{cases}$$

In the RIX population, pairs of RIX sharing one parent are more closely related than those RIX that do not share a parent. For example, RIX produced by crossing RI_1 and RI_2 (RIX_{12}) is expected to be more similar to RIX produced by crossing RI_1 and RI_3 (RIX_{13}) than to RIX from crosses between RI_3 by RI_4 (RIX_{34}) since (RIX_{12}) and (RIX_{13}) share a parental RI (RI_1) while (RIX_{12}) and (RIX_{34}) do not share any parental RI lines.

To model the RIX design, we fit a mixed effect model by applying a random effect to model the polygenic effect. For simplicity, a model with only additive effect is considered. Also, we assume that all putative QTL are located on markers, which is reasonable with

a dense map. The model can be expressed as

$$y_i = \mu(t_i) + x_i\beta(t_i) + \sum_{l=0}^{L} a_{il}\alpha_l + \epsilon_i, \tag{2.1}$$

where y_i is the measure of the genotype of individual i; t_i is the value of some covariate for individual i; $\mu(t_i)$ is the overall effect of the covariate; x_i is the genotype of the ith individual at a certain marker, coded as -1, 0 or 1 for genotypes aa, Aa and AA, respectively; $\beta(t_i)$ is the QTL effect for the covariate t_i; the random polygenic effect α_l follows $N(0, \sigma_a^2)$ for $l = 1, 2, ..., L$; the random error ϵ_i follows $N(0, \sigma_0^2)$; and

$$a_{il} = \begin{cases} 1, & \text{if one of } ith \text{ individual's parents is } RI_l; \\ 0, & \text{otherwise .} \end{cases}$$

The hypotheses for whether there exists any major QTL at a given locus are $H_0 : \beta(t) = 0$ vs $H_a : \beta(t) \neq 0$.

We incorporate B-spline approximation to model the functional QTL effect $\beta(t)$. The smoothness of the function modeled by B-splines is controlled by the smoothness parameter $K = n_j + d + 1$, where n_j is the number of interior knots and d is the degree of splines. Increasing n_j or d will enhance the maximum number of pieces, or the order of polynomials, for piecewise polynomials (splines), respectively, and hence improve the smoothness of the function as the linear combination of B-splines, $\beta(t) = \sum_{k=1}^{K} \gamma_k B_k(t)$, where $B_k(t)$'s are basis functions. Basis functions can be iteratively generated by

$$B_{k,0}(t) = \begin{cases} 1, & \text{if } t_k \leq t < t_{k+1} \text{ ;} \\ 0, & \text{otherwise .} \end{cases}$$

For $c = 1, ..., d$ and $k = 1, ..., K$

$$B_{k,c}(t) = \frac{t - t_k}{t_{k+c} - t_k} B_{k,c-1}(t) + \frac{t_{k+c+1} - t}{t_{k+c+1} - t_{k+1}} B_{k+1,c-1}(t),$$

until $B_{k,d}(t)$'s are obtained as basis B-splines, which are denoted as $B_k(t)$'s. The interior knots of the splines can be either equally spaced or placed on the sample quantiles of the data, so that there are about the same number of observations between any two adjacent knots. We will use equally spaced knots for all numerical examples for this study, and hence $B_k(t)$ is predetermined for any given t.

The mixed effects model becomes

$$y_i = \sum_{k=1}^{K} \gamma_{0k} B_k(t_i) + \sum_{k=1}^{K} \gamma_k B_k(t_i) x_i + \sum_{l=0}^{L} a_{il}\alpha_l + \epsilon_i,$$

where $B_k(t_i)$'s are basis functions of B-splines of order K, γ_k's are coefficients for B-spline basis, and the intercept $\mu(t_i)$ is modeled as $\sum_{k=1}^{K} \gamma_{0k}(t_i) B_k(t_i)$, similar to $\beta(t_i)$. The hypotheses are then equivalent to $H_0 : \gamma_1 = ... = \gamma_K = 0$ vs $H_a : \gamma_1 \neq 0, ..., \gamma_K \neq 0$.

We can rewrite the model above into matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{A}\boldsymbol{\alpha} + \boldsymbol{\epsilon},$$

where $\mathbf{y} = (y_1, ...y_n)^T$; $\boldsymbol{\gamma} = (\gamma_{01}...\gamma_{0K}, \gamma_1...\gamma_K)^T$; $\mathbf{X}$ is the corresponding $n \times 2K$ design matrix for the fixed effect; $\boldsymbol{\alpha} = (\alpha_1, ...\alpha_L)^T$; $\boldsymbol{\epsilon} = (\epsilon_1, ...\epsilon_n)^T$; and $\mathbf{A} = a_{il}$ is an $n \times L$ design matrix for the random polygenic effect. The design matrix $\mathbf{X}$ can be expressed as

$$\mathbf{X} = \begin{pmatrix} B_1(t_1) & \dots & B_K(t_1) & x_1 B_1(t_1) & \dots & x_1 B_K(t_1) \\ \vdots & \vdots & \vdots & \vdots & & \\ B_1(t_n) & \dots & B_K(t_n) & x_n B_1(t_n) & \dots & x_n B_K(t_n) \end{pmatrix}.$$

Therefore, $\mathbf{y}$ follows $N(\mathbf{X}\boldsymbol{\gamma}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \sigma_a^2 \mathbf{A}\mathbf{A}^T + \sigma_0^2 \mathbf{I}$, which can be reparameterized as $\boldsymbol{\Sigma} = \sigma_0^2(\theta\mathbf{D} + \mathbf{I}) = \sigma_0^2 \mathbf{V}$, with $\theta = \frac{\sigma_a^2}{\sigma_0^2}$, $\mathbf{D} = \mathbf{A}\mathbf{A}^T$ and $\mathbf{V} = \theta\mathbf{D} + \mathbf{I}$.

To get an estimator of $\boldsymbol{\gamma}$, the least squares method is not applicable here since the covariance matrix $\boldsymbol{\Sigma}$ is not a diagonal matrix. The generalized least squares (GLS) is

more appropriate

$$\hat{\boldsymbol{\gamma}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}.$$

However, it requires the knowledge of $\boldsymbol{\Sigma}$. So we need to estimate the parameters in $\boldsymbol{\Sigma}$, σ_0^2 and θ, which can be solved by likelihood-based methods.

The log-likelihood functions, based on maximum likelihood (ML) and restricted/residual maximum likelihood (REML), can be written as

$$-2l(\sigma_0^2, \theta|y) = log|\mathbf{V}| + nlog(\sigma_0^2) + \sigma_0^{-2}\mathbf{r}^T\mathbf{V}^{-1}\mathbf{r} + nlog(2\pi),$$

for ML and

$$-2l_R(\sigma_0^2, \theta|y) = log|\mathbf{V}| + (n-p)log(\sigma_0^2) + log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}| + \sigma_0^{-2}\mathbf{r}^T\mathbf{V}^{-1}\mathbf{r} + (n-p)log(2\pi)$$

for REML, where $\mathbf{r} = \mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ and p is the rank of $\mathbf{X}$. This profiles out the parameter $\boldsymbol{\gamma}$ and proves an objective function for σ_0^2 and θ.

However, it is challenging to directly solve for the ML or REML estimate of both σ_0^2 and θ. So we profile out σ_0^2 out of log-likelihood functions (Wolfinger et al. 1994) by expressing it a function of $\mathbf{r}$ and θ,

$$\hat{\sigma_0^2} = \frac{1}{n}\mathbf{r}^T\mathbf{V}^{-1}\mathbf{r},$$

for ML and

$$\hat{\sigma_0^2} = \frac{1}{n-p}\mathbf{r}^T\mathbf{V}^{-1}\mathbf{r}$$

for REML. Substitute the expressions above, we obtain the profile log-likelihoods for θ as

$$-2l(\theta|y) = log|\mathbf{V}| + nlog(\mathbf{r}^T\mathbf{V}^{-1}\mathbf{r}) + nlog(2\pi),$$

and

$$-2l_R(\theta|y) = log|\mathbf{V}| + log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}| + (n-p)log(\mathbf{r}^T\mathbf{V}^{-1}\mathbf{r}) + (n-p)log(2\pi).$$

Note that the profile log-likelihood above only involves the nuisance parameter θ. Hence its MLE can be easily computed by Newton-Raphson algorithm. Then $\boldsymbol{\gamma}$ and σ_0^2 can be estimated by

$$\hat{\boldsymbol{\gamma}} = (\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y},$$

and

$$\hat{\sigma_0^2} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\gamma}})^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\gamma}}),$$

for ML and

$$\hat{\sigma_0^2} = \frac{1}{n-p}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\gamma}})^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\gamma}}),$$

for REML. We use REML in the following simulation studies, since it has some advantages over ML, such as taking into account the degrees of freedom for fixed effects (McCulloch and Searle, 2001).

Once the parameters are estimated, likelihood ratio (LR) test can be performed to evaluate the evidence of QTL effect, and LOD scores can be calculated at the locations of all genetic markers

$$LOD = log_{10}L_R(\hat{\boldsymbol{\gamma}}, \hat{\theta}, \hat{\sigma_0^2}) - log_{10}L_R(0, \tilde{\theta}, \tilde{\sigma_0^2}),$$

where $(\tilde{\theta}, \tilde{\sigma_0^2})$ is the MLE under $H_0 : \gamma_1 = ... = \gamma_K = 0$.

Since the hypothesis testing is performed on a number of markers, it is necessary to adjust the significance level for multiple testing and the maximum LOD should be compared to certain threshold to access the significance of hypothesis testing. The permutation procedure, introduced by Churchill and Doerge (1994), is the most commonly used method to determine the threshold values. It generates permuted samples of the

data by randomly pairing marker genotypes with phenotypes, where the threshold value can be obtained as the $100(1-\alpha)$th percentile of the empirical distribution of test statistics, such as LOD scores, created from permuted samples. However, for RIX data, direct permutation destroys not only the relationship between the major QTL and the trait, but also the relationship between polygenes and the trait, which will result in lower threshold than the true value (Zou et al. 2005). To overcome this difficulty, we apply the nested permutation method (Zou et al. 2005) to RIX data, which permutes genotypes of parental RI stains and creates new marker genotypes of RIX corresponding to the permuted RI stains. The permuted samples are analyzed with the same model as the original data to generate a set of permuted LOD scores where the threshold value is obtained.

2.3 Results

In simulation studies, we applied the loop design for mating scheme as described by Zou et al. (2005), and we set number of RIX lines $L = 100$ and the number of subjects $n = 300$. A single chromosome with 101 evenly spaced markers are simulated with either a 2cM-interval or 5cM-interval between nearby markers (resulting in a total length of 2 Morgan or 5 Morgan). The QTL is located at the 41th marker, which is either at 80 cM or 200 cM. The marker genotypes are simulated using R/qtl (Broman et al. 2003). We assume that the mean temporal growth pattern for QTL genotype Aa is $\mu(t) = \frac{10}{1+5e^{-0.1t}}$, a logistic growth curve (Ma et. al 2002, Yang et al. 2009), where t is randomly generates from $(0, 60)$ for each subject.

We considered the 3 different functions for the functional genetic effect $\beta(t)$, with the range of t being $0 < t < 60$.

Case 1: $\beta(t) = 1 + 3sin(\frac{\pi t}{30})$.

Case 2: $\beta(t) = 1 + \frac{(30-t)^3}{5000}$.

Case 3: $\beta(t) = \frac{3}{2}(arctan(\frac{t-30}{4}) + \frac{\pi}{2})$.

Case 1, a periodical functional effect, and case 2, a non-linearly increasing functional effect, are used in simulation studies by Huang et al. (2008). Case 3 is to mimic the situation of some gene whose effect does not show up until certain age, such as some breast cancer-susceptibility genes (Foulkes et. al 2004). To test the performance of the model under various signal/noise ratios, two different sets of variances for random effect and random error are considered for each case: $\sigma_a^2 = 10$, $\sigma_0^2 = 20$ and $\sigma_a^2 = 30$, $\sigma_0^2 = 30$. In all cases, the average heritability is between 0.02 and 0.18.

To choose a good combination of interior knots number n_j and degree of spline d to model the genetic effect, we first used a set of combinations to fit the true model with the QTL genotype and the intercept as predictors. 500 runs of simulation were performed and the mean estimated $\mu(t)$ and $\beta(t)$ are computed. In those simulations, we set $\sigma_a^2 = 30$, $\sigma_0^2 = 30$ and 5 cM intervals between markers. Figures 2.1-2.4 plotted the mean $\hat{\mu}(t)$ in case 1 and mean $\hat{\beta}(t)$ in all cases for each combination of n_j and d, which showed that relatively small number of n_j and d is enough to fit the curves well. We calculated the squared differences (SQD) between $\hat{\mu}(t)$ and $\mu(t)$, and between $\hat{\beta}(t)$ and $\beta(t)$ by $SQD = \int_{t=0}^{60} \{(\hat{\mu}(t) - \mu(t))^2 + (\hat{\beta}(t) - \beta(t))^2\} dt$ for each choice of n_j and d. We counted how many times that any combination of n_j and d has the smallest SQD and recorded the counts in the right panel of table 2.1. The results indicates that $n_j = 1$ and $d = 2$ is the best choice for cases 1 and 2; and $n_j = 2$ and $d = 1$ is the best choice for case 3.

In practice, the true $\beta(t)$ is unknown, so the choice of n_j and d needs to be decided by the data. AIC (Akaike 1970, 1974) can be used as a criterion to select a reasonable degree and smooth of B-splines. The model with smaller AIC is superior. We propose the following approach to choose n_j and d. First, set $n_j = 1$ and $d = 1$, identify the marker with the highest LOD score. Then calculate the AIC values for the marker picked in the previous step for a set of values of n_j and d, and choose the combination of n_j and d with the smallest AIC. In the simulation study, we computed the AIC values for the

500 runs of simulations. The left panel of table 2.1 recorded the number of runs where each combination of n_j and d has the smallest AIC. The results are close to the right part of the table that is determined by SQD.

To compare with the nonparametric model, we also fitted the varying coefficient with parametric models. We used polynomial functions to model $\beta(t)$

$$\beta(t) = \sum_{k=0}^{s} \gamma_k t^k,$$

where s is the order of polynomials. We set $s = 1$ and $s = 2$, for linear and quadratic polynomial functions, in the simulation studies.

Under each case, 200 runs of simulations were conducted with all models mentioned above. For each case, parameters have been estimated using the method described above. From that, we compute the estimated $\beta(t)$ for B-splines and polynomials. Hypothesis testing have been perform on $H_0 : \beta(t) = 0$ vs $H_a : \beta(t) \neq 0$, and LOD scores are calculated from likelihood ratio tests. To get the threshold values for access the significance of hypothesis testing, simulations were carried out using the model that involve no genetic effect

$$y_i = \mu(t_i) + \sum_{l=0}^{L} a_{il}\alpha_l + \epsilon_i.$$

Likelihood ratio tests were conducted on each marker and the maximum LOD score was recorded for each run of simulation. 1000 runs of simulations was performed and the 95% percentile of the maximum LOD score is used as a cutoff value to access the significance of likelihood ratio tests.

In practice, it is impossible to calculate the empirical threshold of LOD score, since the random errors are unknown. Hence, we need to obtain the permuted threshold of LOD score by the nested permutation. To evaluate the performance of the nest permutation method, we carried out the follow simulation studies. We simulated 300 subjects from 100 RIX lines. A single 100cM chromosome with evenly spaced markers are

simulated and the QTL is located at the 40 cM. There are either 51 markers separated by 2cM-intervals or 21 markers separated by 5cM-intervals, on the 100cM chromosome. 2 different functions, case 1 and case 2 as described above, are considered for the varying coefficient $\beta(t)$. We set $\mu = 0$, $\sigma_a^2 = 30$, and $\sigma_0^2 = 30$. 50 runs of simulation are conducted, each with 1000 nested permutation samples. The 95% percentile of the maximum LOD values computed for each run of simulation as the permutation threshold. Powers of LR tests are computed by comparing the maximum LOD scores to the threshold values. The results, listed in table 2.3, indicated that the nested permutation performs well because the permutation threshold values are very close to the empirical threshold values and are not affected by the choice of $\beta(t)$. Also, the powers under $H_0 : \beta(t) = 0$ are very close to the significance level, 0.05.

The location of the QTL is estimated as the location where the maximum LOD is reached. The means and standard errors of the estimated genetic location of the QTL, by the three approaches, are listed in table 2.3. Powers of hypothesis testings are listed in table 2.4. All the three methods provide similar estimations of QTL location and powers for QTL detection under cases 2 and 3. However, the B-spline approach produces substantially higher power than other two approaches in case 1, as well as higher precision in estimating the QTL location. As expected, more power and more precise estimates can be obtained with smaller variances for each approach. The estimated phenotypic mean curves $E\{y(t)\} = \hat{\mu}(t) + x\hat{\beta}(t)$ are plotted along time in figures 2.5-2.7, for all cases with 5 cM intervals, $\sigma_a^2 = 10$ and $\sigma_0^2 = 20$. The nonparametric approach provides better fit to the true underlying phenotypic mean curves than the parametric approach. In all three cases, the estimated curves by the B-spline approach generally has less deviation from the true curves for all three genotypes, comparing to the estimated curves by the parametric approach. Overall speaking, the B-spline method outperforms the parametric method.

2.4 Discussion

Recombinant inbred intercrosses process some good properties from both RI lines and F2 populations. Genotypes of RIX can be directly inferred from those of their parental RI lines. Unlike the parental RIs whose genotypes are homozygous, the genetic structure of an RIX resembles F2 animals, reducing the phenotypic anomalies associated with inbred genomes. One big advantage on using RIX mice for QTL mapping is from the ongoing Collaborative Cross project (Threadgill et al. 2002). The CC project aims to generate and maintain about 1000 multi-parental CC RI lines, and our ability to map complex traits will be greatly increased by making use of those huge amount of resources.

In our simulations, we assume no maternal or paternal effects and thus only non-reciprocal RIX are simulated. If maternal or paternal effects are of interest, reciprocal RIX can be easily generated, and the model can be simply adapted by adding one random effect. Although our model only considers the additive genetic effect, the dominant effect can be easily included in the model by adding one fixed effect. We applied single marker analysis in our simulations because the high marker density of the parental RI, and thus RIX, makes results similar to those that would be obtained using more complicated mapping methods, such as traditional interval mapping (Lander and Botstein 1989) or regression interval mapping (Haley and Knott 1992).

Besides B-spline approximation, other nonparametric approaches can be used to model the varying coefficients, such as the local polynomial regression (Fan and Gijbels 1996), the smoothing splines (Hastie and Tibshirani 1993; Hoover et al. 1998) and wavelet-based approaches (Donoho and Johnstone 1994). One advantage of using B-splines is that the smoothing matrix $\{B_k(t_i)\}$ is independent of the responses, so that empirical threshold values for nonparametric functional mapping can be obtained through simulations. Unlike other nonparametric approaches, how to determine the smoothness is still an open question, although the choice of the number of knots is generally not critical (Yang et al. 2009). Our simulation results (figures 1.1-1.4) show that

the estimated functional effects are not sensitive to the choices of d and n_j.

When there is more than one QTL on a chromosome, the test statistic at one position will be affected by all those QTL, the estimates are likely to be biased, and QTL can be mapped to wrong positions (Knott and Haley 1992; Martinez and Curnow 1992). Single marker analysis cannot tell whether the markers are associated with one or more QTL. Our model can be extended to multiple regression for multiple QTL mapping, and some model selection approaches can be applied to selection important genes, as will be discussed in details in the next chapter.

		AIC					SQD				
	$n_j =$	1	2	3	4	5	1	2	3	4	5
Case 1	$d = 1$	40	126	36	21	3	0	73	39	10	0
	$d = 2$	182	14	9	4	3	326	19	4	1	1
	$d = 3$	28	4	2	2	1	17	2	0	0	0
	$d = 4$	7	8	5	2	3	6	1	0	1	0
Case 2	$d = 1$	122	85	28	18	6	47	73	32	7	0
	$d = 2$	155	17	12	4	2	257	34	5	1	1
	$d = 3$	19	2	2	2	3	36	2	0	0	0
	$d = 4$	7	7	4	2	3	3	1	0	1	0
Case 3	$d = 1$	136	180	12	24	6	68	259	13	18	1
	$d = 2$	78	6	10	2	4	114	4	6	1	1
	$d = 3$	13	2	1	2	3	9	1	0	0	0
	$d = 4$	8	4	5	1	3	4	0	0	1	0

Table 2.1: Counts of selections by the smallest AIC or SQD

		$\beta(t) = 0$	Case1	Case2	Empirical
LOD score	2 cM	3.5165	3.5210	3.5171	3.5274
	5 cM	3.7973	3.7876	3.8123	3.8198
Power	2 cM	0.06	0.74	0.70	0.05
	5 cM	0.06	0.82	0.60	0.05

Table 2.2: Mean threshold LOD and power by nest permutation and simulations

Variance	Distance	Method	Case1	Case2	Case3
$\sigma_a^2 = 10$, $\sigma_0^2 = 20$	2 cM	B-spline	79.59 (0.71)	81.60 (1.17)	80.80 (0.74)
		Linear	79.46 (1.55)	83.62 (1.19)	81.01 (0.74)
		Polynomial	80.23 (1.41)	83.13 (1.55)	82.66 (1.10)
	5 cM	B-spline	198.15 (2.11)	203.28 (3.26)	201.75 (1.43)
		Linear	203.15 (3.74)	209.00 (3.29)	201.05 (1.25)
		Polynomial	209.80 (5.08)	211.98 (4.43)	204.95 (2.37)
$\sigma_a^2 = 30$, $\sigma_0^2 = 30$	2 cM	B-spline	80.02 (1.93)	81.92 (2.15)	81.78 (1.32)
		Linear	79.81 (2.65)	84.65 (1.99)	81.88 (1.20)
		Polynomial	79.88 (2.56)	83.63 (2.21)	82.42 (1.38)
	5 cM	B-spline	204.03 (5.73)	207.20 (5.95)	206.10 (4.40)
		Linear	210.23 (6.74)	209.68 (5.47)	203.80 (3.19)
		Polynomial	201.08 (7.07)	207.70 (5.76)	208.33 (4.25)

Table 2.3: The mean and standard error of estimated QTL location

Variance	Distance	Method	Case1	Case2	Case3
$\sigma_a^2 = 10$, $\sigma_0^2 = 20$	2 cM	B-spline	0.855	0.775	0.930
		Linear	0.620	0.735	0.920
		Polynomial	0.600	0.695	0.895
	5 cM	B-spline	0.845	0.735	0.930
		Linear	0.590	0.725	0.905
		Polynomial	0.590	0.695	0.875
$\sigma_a^2 = 30$, $\sigma_0^2 = 30$	2 cM	B-spline	0.490	0.430	0.660
		Linear	0.260	0.430	0.650
		Polynomial	0.245	0.385	0.625
	5 cM	B-spline	0.450	0.450	0.695
		Linear	0.220	0.395	0.685
		Polynomial	0.195	0.360	0.670

Table 2.4: Power of likelihood ratio test for the three approaches

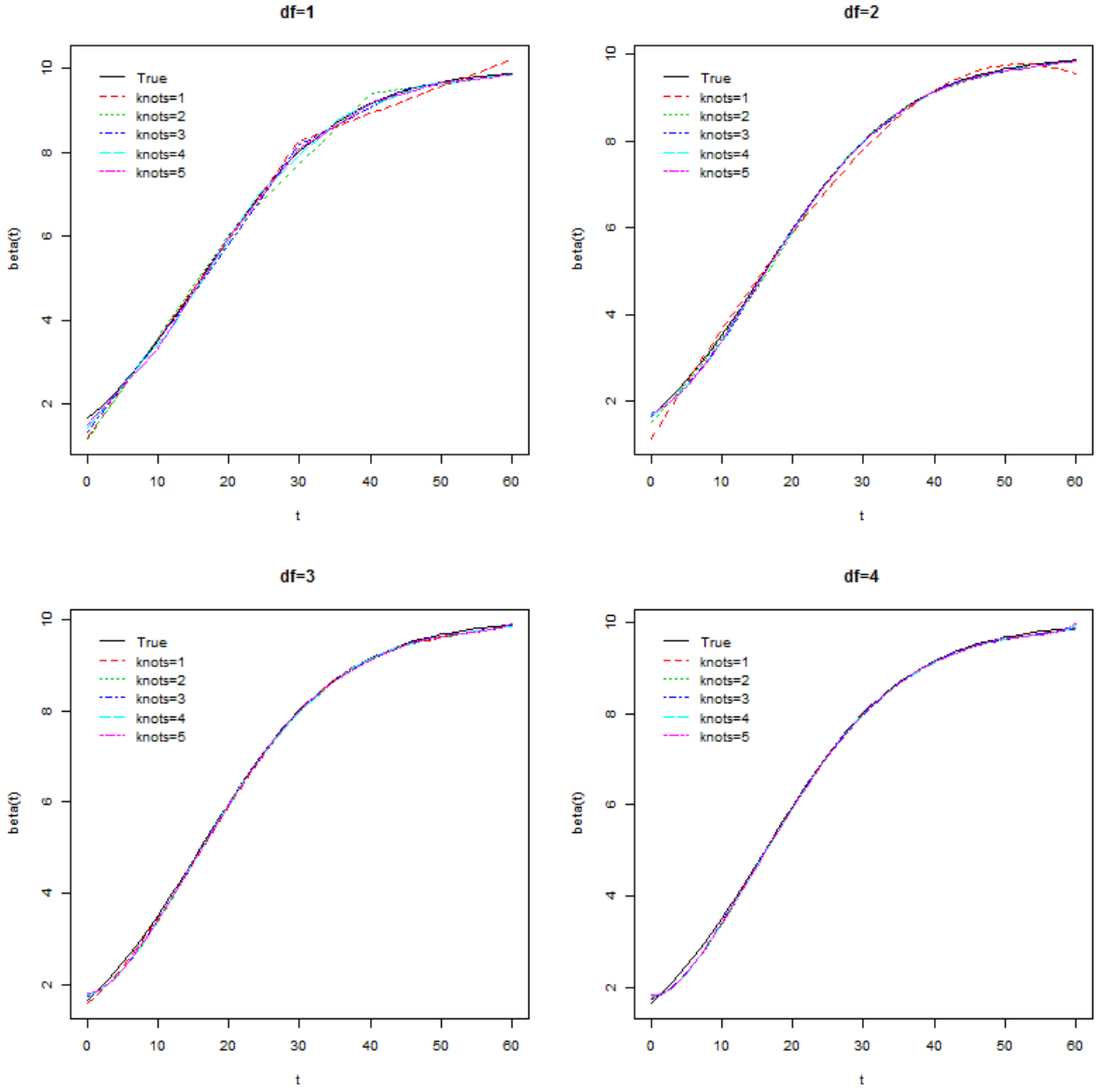

Figure 2.1: The true and estimated curves for $\mu(t) = \frac{10}{1+5e^{-0.1t}}$

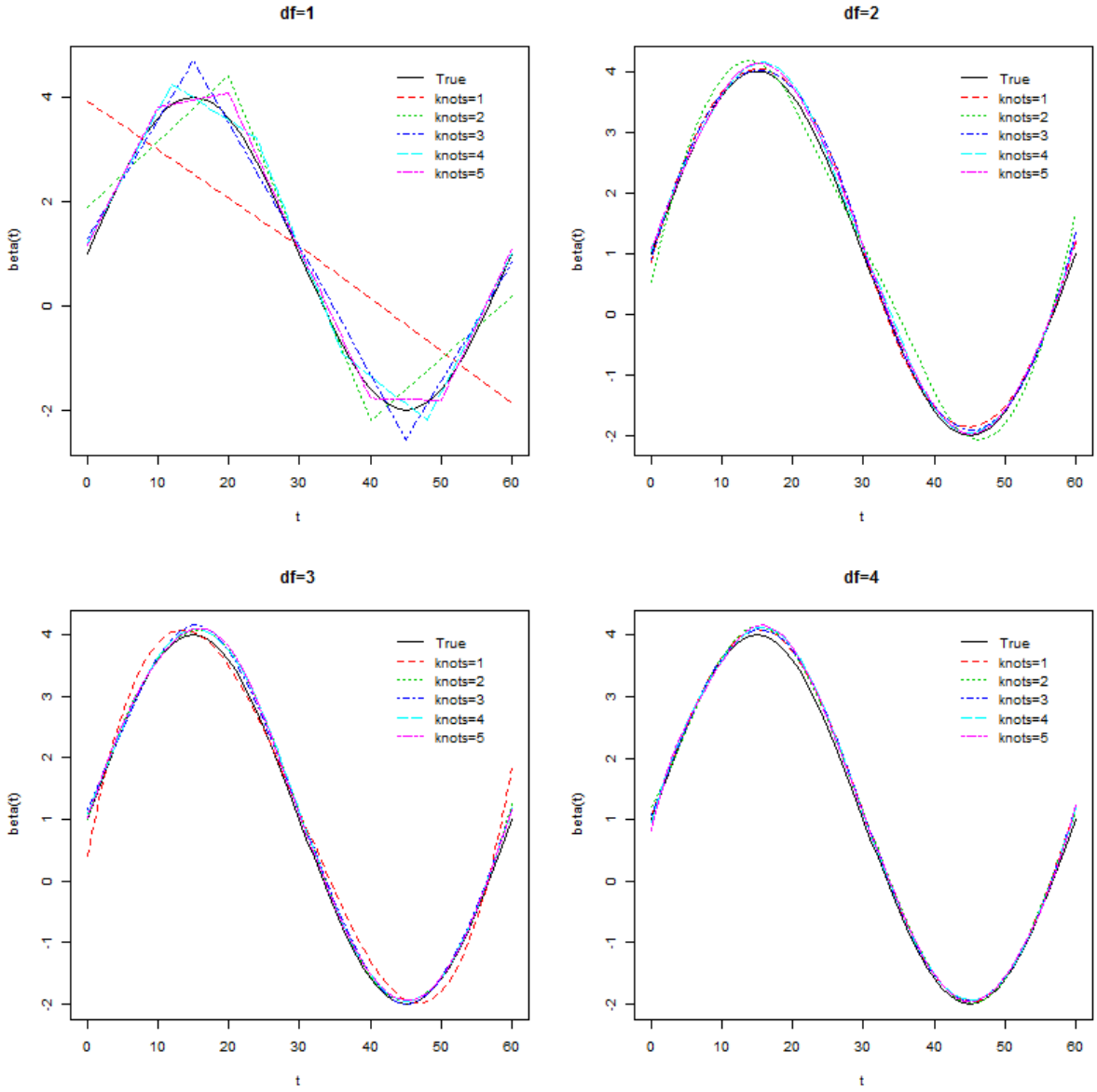

Figure 2.2: The true and estimated curves for $\beta(t) = 1 + 3sin(\frac{\pi t}{30})$

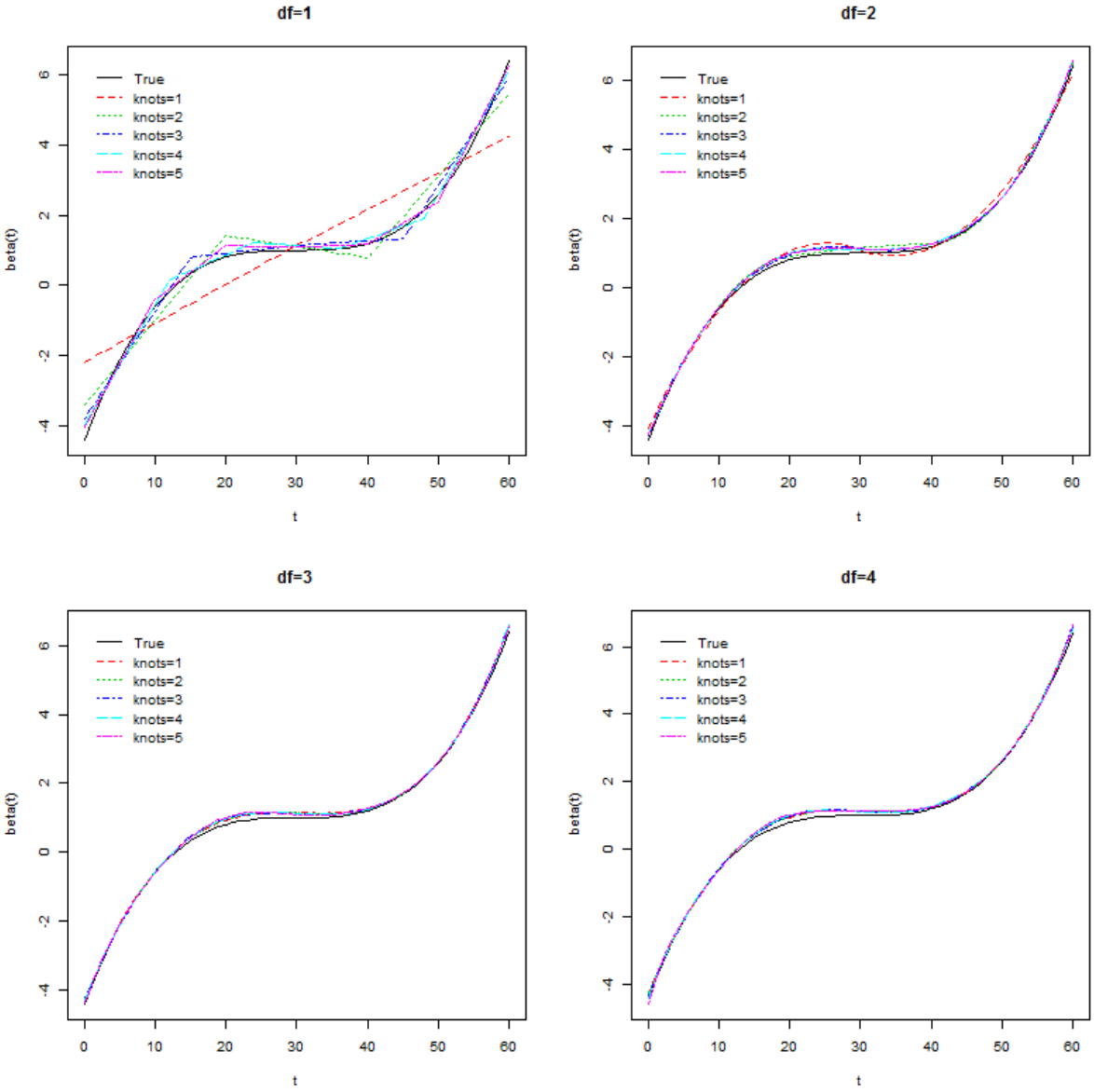

Figure 2.3: The true and estimated curves for $\beta(t) = 1 + \frac{(30-t)^3}{5000}$

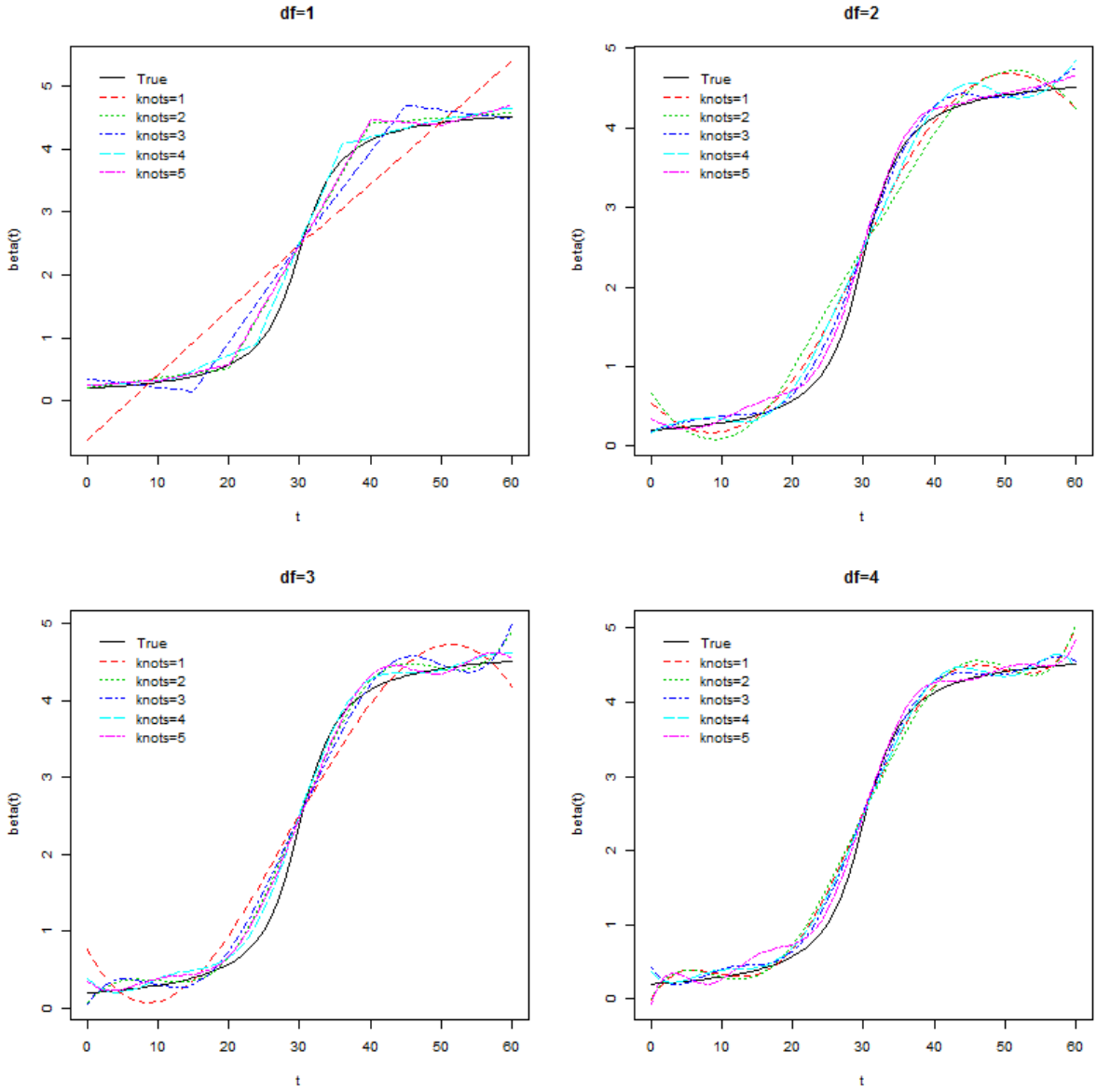

Figure 2.4: The true and estimated curves for $\beta(t) = \frac{3}{2}(arctan(\frac{t-30}{4}) + \frac{\pi}{2})$

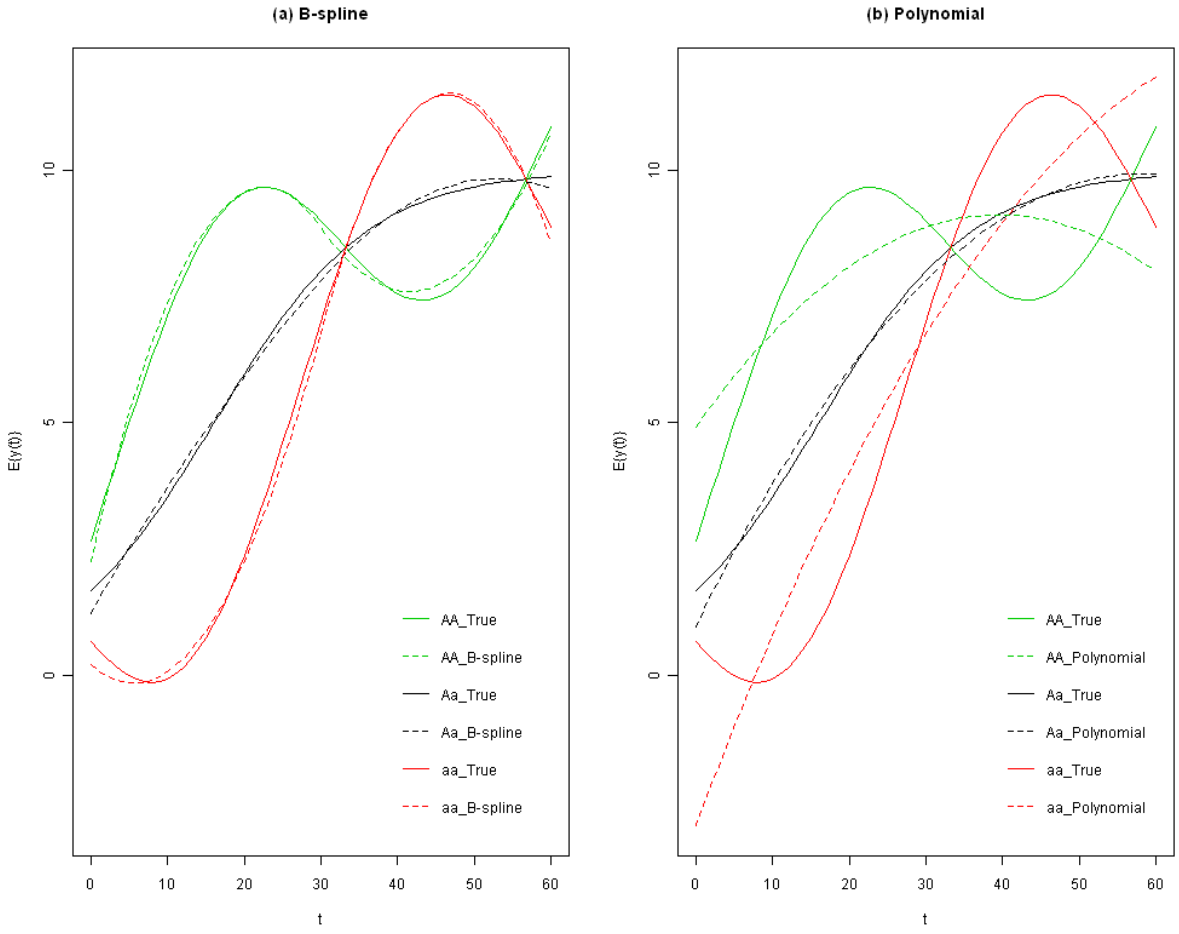

Figure 2.5: $\beta(t) = 1 + 3sin(\frac{\pi t}{30})$ estimated by (a) B-spline and (b) Polynomial

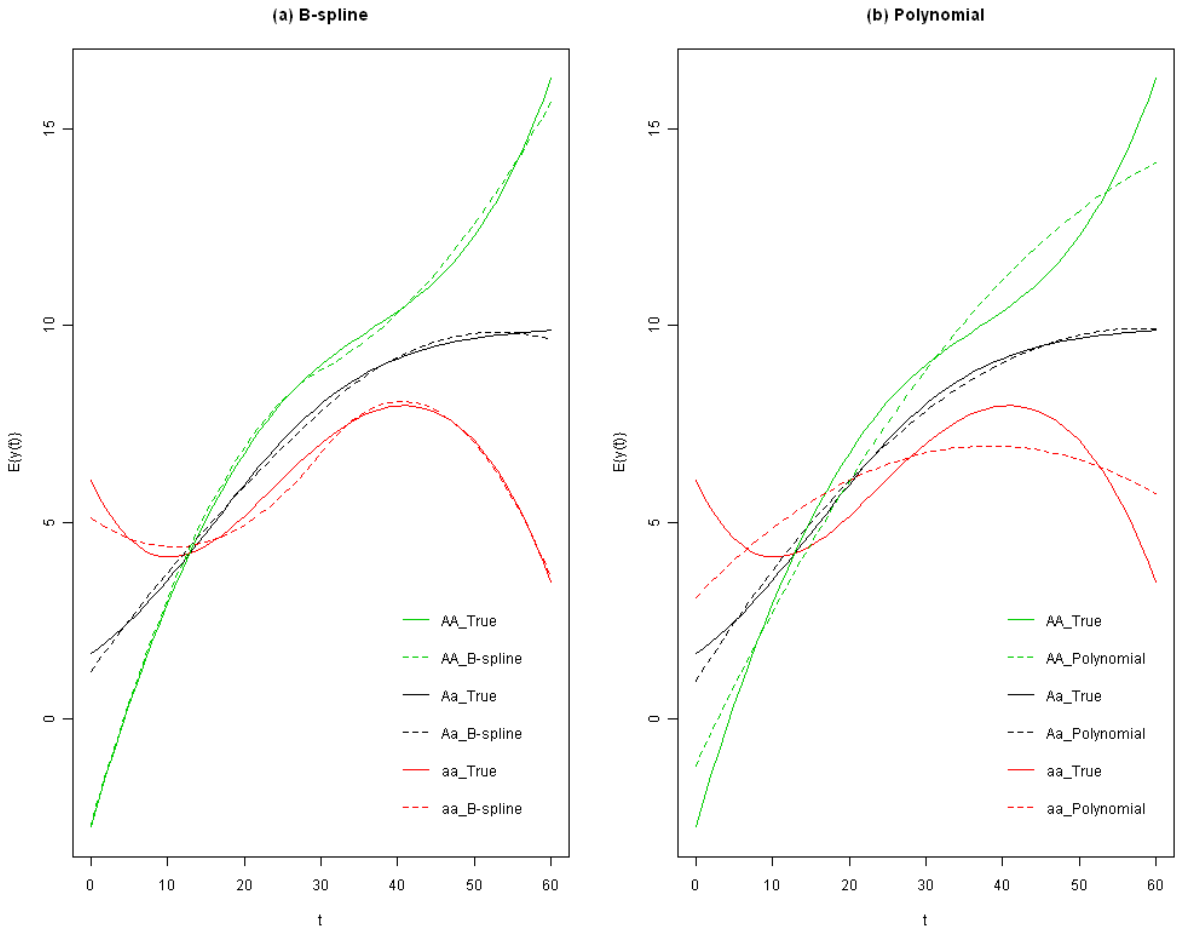

Figure 2.6: $\beta(t) = \frac{5}{1+e^{-0.1t}}$ estimated by (a) B-spline and (b) Polynomial

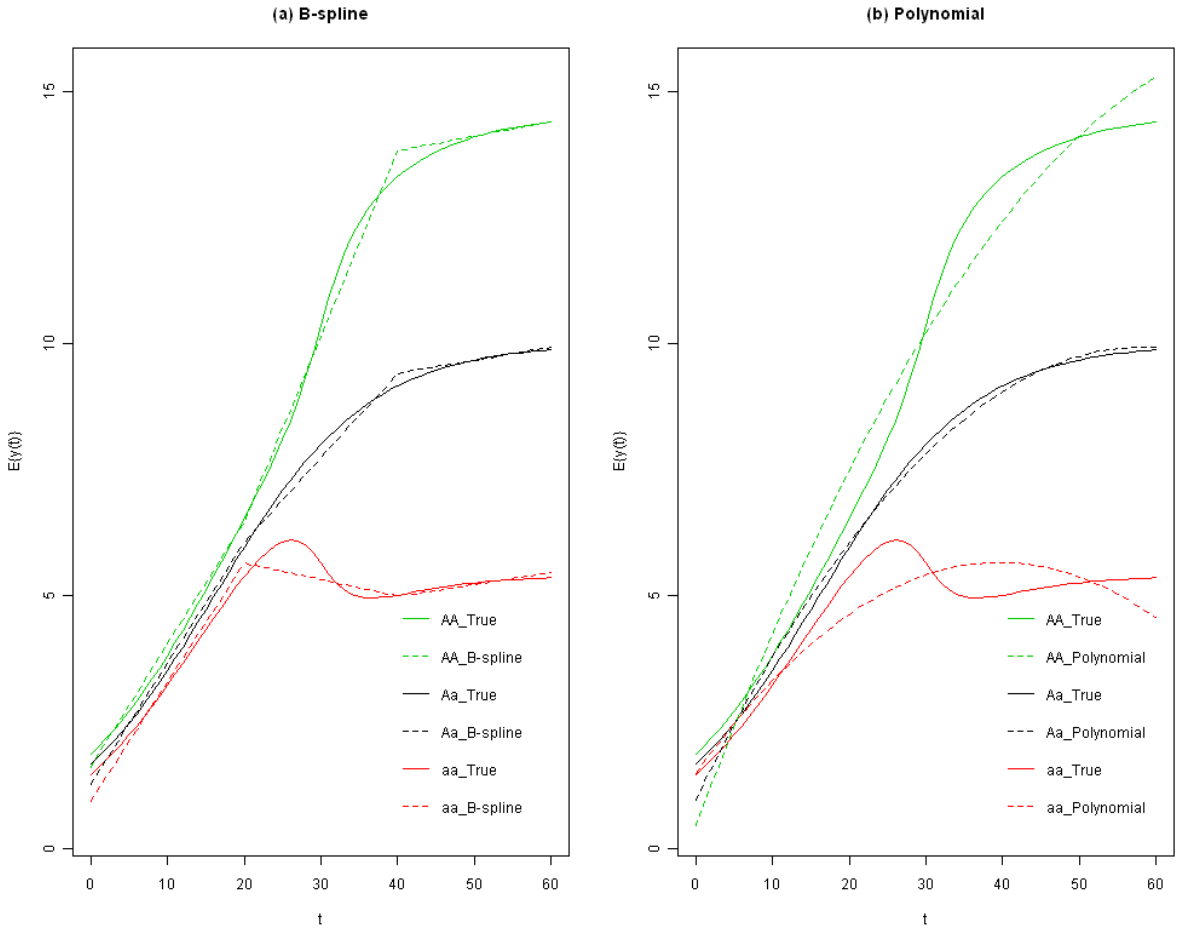

Figure 2.7: $\beta(t) = \frac{3}{2}(tan^{-1}(\frac{t-30}{4}) + \frac{\pi}{2})$ estimated by (a) B-spline and (b) Polynomial

Chapter 3

Varying Coefficient Models for Multiple QTL mapping

3.1 Introduction

In the previous chapter, we assumed one QTL model. However, a large amount of traits in nature are affected by many genes (Zeng 1994). When there is more than one QTL on a chromosome, the test statistic at one position, derived from one QTL model, will be affected by all those QTL, the estimates are likely to be biased, and QTL can be mapped to wrong positions (Knott and Haley 1992; Martinez and Curnow 1992). Single marker analysis cannot tell whether the markers are associated with one or more QTL. For example, Wright and Kong (1997) showed that with single-gene models, an apparent "ghost" gene can appear between two true QTL. Therefore, it is desirable to develop any method that is able to map multiple QTL at the same time.

To overcome the problem of single QTL model, Zeng (1993, 1994) and Jansen and Stam (1994) proposed the composite interval mapping by adding other markers as covariates in addition to the single QTL model. Picking suitable number of markers (covariates) is important to CIM as too many covariates may decrease the power of QTL mapping and too little is not enough to control genetic background (Broman and Speed

2002). Furthermore, it is hard to access the genome-wide significance for CIM. Kao and Zeng (1997) proposed multiple interval mapping by simultaneously including multiple putative QTL and their interactions in one model. However, the evaluation of MIM model is very computationally intensive due to the large number of unknown parameters (Zeng et al. 2000). Alternatively, QTL mapping can be viewed as one problem of model selection, and model selection procedures, such as stepwise regression, can be used to search for multiple QTL (Broman and Speed 2002). Another type of multiple QTL mapping method is Bayesian QTL mapping (Yi and Xu 2000), which treats the number of QTL as a random variable and models it using reversible jump Markov chain Monte Carlo (Green 1995).

Numerous evidences have displayed that genes have quite different effects on a phenotypic trait of an individual at different ages or different environment (Pletcher and Geyer 1999; Ma et al. 2002). The functional pattern of QTL effect can be implemented with varying coefficient models, where the functional effect can be approximated by B-spline basis. The varying coefficient model in the previous chapter can be extended to multiple QTL mapping by adding multiple markers into the model.

To get the estimation of parameters for varying coefficient models, Hastie and Tibshirani (1993) suggested the penalized least squares method. The most popular penalized least squares method for the linear models is the least absolute shrinkage and selection operator (LASSO) introduced by Tibshirani (1996), which estimates the regression coefficients by minimizing a penalized least squares criterion with L_1 penalty function of regression coefficients. LASSO is able to produce sparse solutions, and thus select a parsimonious model. Tibshirani (1996) also provided a computational algorithm to get LASSO estimates by solving the constrained least squared problem. Fu (1998) developed a "shooting algorithm" for LASSO. Efron et al. (2004) proposed a new model selection algorithm, the least angel regression (LARS), and showed that it turns to LASSO by some simple modification. The LARS algorithm simplified the implementation of LASSO.

LASSO can be applied to high dimensional data, but it cannot select more variables than the number of predictors. LASSO estimates are biased by shrinking toward zero. Furthermore, LASSO may not perform well for the data where predictors have very high correlation. Alternatively, Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) method, which makes use of a non-convex penalty function that does not penalize big β_j. To handel the optimization of the non-convex function, Fan and Li (2001) proposed a unified algorithm for the minimization of penalized likelihood via local quadratic approximations.

In the varying coefficient model above, it will be desirable to select functions of one varying coefficient as a group. To solve the problem of selecting grouped variables, Yuan and Lin (2006) proposed the group LASSO method. The group LASSO estimator is obtained by

$$\hat{\boldsymbol{\beta}} = argmin\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\boldsymbol{\beta}_j x_{ij})^2 + \lambda\sum_{j=1}^{J}(\boldsymbol{\beta}_j^T\mathbf{K}_j\boldsymbol{\beta}_j)^{1/2}\},$$

where $\boldsymbol{\beta}$ consist of J groups of regression coefficients, $\boldsymbol{\beta}_j$. $\mathbf{K}_j$'s are some positive definite kernel matrices, for which a simple choice will be $\mathbf{K}_j$=$\mathbf{I}_{p_j}$, where $\mathbf{I}_{p_j}$ is the p_j-dimensional identity matrix with p_j being the number of regression coefficients in the jth group. Similarly, Yuan and Lin (2006) also extended LARS and non-negative garrote to group LARS and group non-negative garrote. Base on the group LASSO, Lin and Zhang (2006) developed Component Selection and Smoothing Operator (COSSO) for variable selection in smoothing spline ANOVA, and extended the method to nonparametric regression (Zhang and Lin, 2006). Similar to group LASSO, Wang et al. (2007) developed the group SCAD method, and extended it to nonparametric varying-coefficient models for repeated measures (Wang et al. 2008).

To choose the tuning parameter λ, Yuan and Lin (2006) introduced a C_p-type criterion for group LASSO. Empirical evidence suggested that the performance this C_p-type criterion is generally comparable with, and sometimes better than, that of fivefold

cross-validation. Yuan and Lin (2006) showed, through simulation, that these methods outperformed the traditional stepwise backward elimination method.

Another popular method to select the tuning parameter λ, which is more data driven, is the cross-validation. The cross-validation method can be employed to estimate the prediction accuracy of a selected model based on a certain tuning parameter, by fitting the model from the training data set and computer mean squared prediction error from the validation data set using the fitted model. The CV criterion $CV(\lambda)$ is the sum of the mean squared prediction error. The tuning parameter λ can be selected from a predetermined set of values, by minimizing the $CV(\lambda)$. Meier et al. (2008) made use of cross-validation to choose tuning parameters in their group LASSO analysis for logistic regression.

To overcome the deficient that the result depends heavily on how the data is divided, simple cross-validation can be easily extended K-fold cross-validation or, more extremely, leave-one-out cross-validation. A drawback of k-fold CV and leave-one-out CV is that they can be very time consuming in computation. The generalized cross-validation method, first proposed by Craven and Wahba (1979), is an alternative to cross-validation that is faster in computation. Tibshirani (1996) extend the GCV statistic to apply it to the LASSO procedure.

For linear models, GCV is asymptotically equivalent to C_p, AIC, and leave-one-out CV (Shao 1997; Hastie et al. 2001). However, GCV tends to overfit, hence Zou et al. (2004) proposed the BIC-LASSO shrinkage, and they found the BIC criterion is more appropriate, comparing to AIC or C_p, when variable selection is the primary concern. Huang et al. (2009) made use of a slightly different form of BIC-type criterion for his group bridge method and found that tuning based on BIC in general does better than that based on C_p, AIC or GCV in terms of selection at the group and individual variable levels.

Those criteria for choosing tuning parameter, including GCV, C_p, AIC and BIC,

all involve estimating the effective degrees of freedom, $df(\lambda)$, which is an informative measurement of model complexity. For simple linear models, the degrees of freedom is simply the number of predictors in the model. Unfortunately, due to the nonlinear nature of the LASSO, the explicit expression of the degrees of freedom, $df(\lambda)$, is not available. $df(\lambda)$ can be approximately estimated by $df(\lambda) = trace\{\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{W}^-)^{-1}\mathbf{X}^T\}$, as proposed by Tibshirani (1996). Besides that, Zou et al. (2004) stated that the easiest approach is to ignore shrinkage and use $df(\lambda) = q$, where q is the number of non-zero parameters. They also showed that it is an unbiased estimate of $df(\lambda)$ and this approximation is reasonable despite of its simplicity.

The rest of the chapter is organized as follows. In the method section, we developed a penalized likelihood method for multiple functional QTL mapping by group selection of coefficients associated with each gene. In the results section, we presented simulation results for the performance of the model selection method. In the discuss section, we ended the chapter by conclusion and discussion.

3.2 Methods

3.2.1 Model

For multiple mapping of QTL with time-varying effects using the Collaborative Cross (CC) panel data, we simply extend the mixed effect model (2.1) to

$$y_i = \mu(t_i) + \sum_{j=1}^{p} x_{ij}\beta_j(t_i) + \sum_{l=0}^{L} a_{il}\alpha_l + \epsilon_i, \tag{3.1}$$

where y_i is the measure of the genotype of individual i; t_i is the value of some covariate for individual i; $\mu(t_i)$ is the overall effect of the covariate; x_{ij} is the genotype of the ith individual at the jth marker, coded as -1, 0 or 1 for genotypes aa, Aa and AA, respectively; $\beta(t_i)$ is the QTL effect for the covariate t_i; the random polygenic effect α_l

follows $N(0, \sigma_a^2)$ for $l = 1, 2, ..., L$; the random error ϵ_i follows $N(0, \sigma_0^2)$; and

$$a_{il} = \begin{cases} 1, & \text{if one of } ith \text{ individual's parents is } RI_l; \\ 0, & \text{otherwise .} \end{cases}$$

For simplicity, only additive effects are considered in our model, and it can be easily extended to include dominant effects.

We model the varying coefficient using functional approximation with B-spline basis, the model becomes

$$y_i = \mu(t_i) + \sum_{j=1}^{p} \sum_{k=1}^{K} x_{ij} \gamma_{jk} B_{jk}(t_i) + \sum_{l=0}^{L} a_{il} \alpha_l + \epsilon_i,$$

where $B_{jk}(t_i)$'s are basis functions of B-splines of order K and γ_{jk}'s are corresponding coefficients. We can rewrite the model above into matrix form as

$$\mathbf{y} = \mathbf{X}_b \boldsymbol{\gamma} + \mathbf{A}\boldsymbol{\alpha} + \boldsymbol{\epsilon},$$

where $\mathbf{y} = (y_1, ...y_n)^T$; $\mu(t_i) = \sum_{k=1}^{K} \gamma_{0k}(t_i) B_{0k}(t_i)$; $\boldsymbol{\gamma} = (\gamma_{01}...\gamma_{0K}, \gamma_{11}...\gamma_{pK})^T$; $\mathbf{X}_b$ is the corresponding $n \times ((p+1)K)$ design matrix for the time varying fixed effect; $\boldsymbol{\alpha} = (\alpha_1, ...\alpha_L)^T$; $\boldsymbol{\epsilon} = (\epsilon_1, ...\epsilon_n)^T$; and $\mathbf{A} = a_{il}$ is a $n \times L$ design matrix for the random polygenic effect. The design matrix for fixed effect, $\mathbf{X}_b$ can be expressed as

$$\mathbf{X}_b = \begin{pmatrix} B_{01}(t_1) & \dots & B_{0K}(t_1) & \dots & x_{1p}B_{p1}(t_1) & \dots & x_{1p}B_{pK}(t_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B_{01}(t_n) & \dots & B_{0K}(t_n) & \dots & x_{np}B_{p1}(t_n) & \dots & x_{np}B_{pK}(t_n) \end{pmatrix}.$$

Therefore, $\mathbf{y}$ follows $N(\mathbf{X}_b\boldsymbol{\gamma}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \sigma_a^2 \mathbf{A}\mathbf{A}^T + \sigma_0^2 \mathbf{I}$. So the log-likelihood function can be computed as

$$-2l(\boldsymbol{\gamma}, \sigma_a^2, \sigma_0^2) = log(|\boldsymbol{\Sigma}|) + (\mathbf{y} - \mathbf{X}_b\boldsymbol{\gamma})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}_b\boldsymbol{\gamma}).$$

Let $\boldsymbol{\Sigma} = \sigma_0^2(\theta\mathbf{D}+\mathbf{I}) = \sigma_0^2\mathbf{V}$, with $\theta = \frac{\sigma_a^2}{\sigma_0^2}$, $\mathbf{D} = \mathbf{A}\mathbf{A}^T$ and $\mathbf{V} = \theta\mathbf{D}+\mathbf{I}$. The log-likelihood function can be be reparameterized as

$$-2l(\boldsymbol{\gamma}, \sigma_0^2, \theta) = log|\mathbf{V}| + nlog(\sigma_0^2) + \sigma_0^{-2}(\mathbf{y} - \mathbf{X}_b\boldsymbol{\gamma})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}_b\boldsymbol{\gamma}),$$

similarly, the restricted/residual maximum likelihood (REML) can be expressed as

$$-2l_R(\boldsymbol{\gamma}, \sigma_0^2, \theta) = log|\mathbf{V}| + (n-pK)log(\sigma_0^2) + |\mathbf{X}_b^T\mathbf{V}^{-1}\mathbf{X}_b| + \sigma_0^{-2}(\mathbf{y}-\mathbf{X}_b\boldsymbol{\gamma})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}_b\boldsymbol{\gamma}).$$

To perform variable selection, we can impose penalty functions, $p_\lambda(||\boldsymbol{\gamma}_j||)$, similar to the group LASSO, to the log-likelihood function above. Then the objective function can be expressed as

$$F(\boldsymbol{\gamma}, \sigma_a^2, \sigma_0^2) = log|\mathbf{V}|+(n-pK)log(\sigma_0^2)+|\mathbf{X}_b^T\mathbf{V}^{-1}\mathbf{X}_b|+\sigma_0^{-2}(\mathbf{y}-\mathbf{X}_b\boldsymbol{\gamma})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}_b\boldsymbol{\gamma})+\sum_{j=1}^{p} p_\lambda(||\boldsymbol{\gamma}_j||),$$

where $p_\lambda(u) = \lambda u$, $\boldsymbol{\gamma}_j = (\gamma_{j1}, ..., \gamma_{jK})^T$ and $||\boldsymbol{\gamma}_j||$ denotes some L_2-norm of $\boldsymbol{\gamma}_j$. For the choice of penalty function, we applied a modified group LASSO penalty function, similar to the group SCAD by Wang et al. (2008), to accommodate the correlation between bases functions. The penalty function can be expressed as

$$p_\lambda(||\boldsymbol{\gamma}_j||) = \lambda||\boldsymbol{\gamma}_j||,$$

where $||\boldsymbol{\gamma}_j||$ is defined as the L_2-norm of the function $\beta_j(t) = \sum_{k=1}^{K}\gamma_{jk}B_{jk}(t)$. The squared L_2-norm can be written as quadratic form $||\boldsymbol{\gamma}_j||^2 = \boldsymbol{\gamma}_j^T\mathbf{R}_j\boldsymbol{\gamma}_j$, where $\mathbf{R}_j = (r_{uv})_{K\times K}$ is a matrix with entries $r_{uv} = \int_{t_1}^{t_2} B_{ju}(t)B_{jv}(t)dt$ with (t_1, t_2) being the range of t.

3.2.2 Computational Algorithm

Fan and Li (2001) proposed the local quadratic approximations method as a unified algorithm to solve for the penalized likelihood problems, where the penalty function $p_\lambda(|\beta_j|)$ can be locally approximated around some initial estimate β_{j0}. If β_{j0} is very close to 0, then set $\hat{\beta_{j0}} = 0$. Otherwise, $p_\lambda(|\beta_j|)$ can be approximated as in formula (1.1).

This approximation can be applied to the penalty we used. Given an initial value of $\boldsymbol{\gamma}^0$ for $\boldsymbol{\gamma}$, $p_\lambda(||\boldsymbol{\gamma}_j||)$ can be approximated by

$$p_\lambda(||\boldsymbol{\gamma}_j||) \approx \frac{1}{2}\frac{p'_\lambda(||\boldsymbol{\gamma}_j^0||)}{||\boldsymbol{\gamma}_j^0||}\boldsymbol{\gamma}_j^T\mathbf{R}_j\boldsymbol{\gamma}_j + Constant.$$

Thus, the objective function can be approximately expressed as

$$F(\boldsymbol{\gamma}, \sigma_0^2, \theta) = log|\mathbf{V}| + (n-pK)log(\sigma_0^2) + |\mathbf{X}_b^T\mathbf{V}^{-1}\mathbf{X}_b| + \sigma_0^{-2}(\mathbf{y}-\mathbf{X}_b\boldsymbol{\gamma})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}_b\boldsymbol{\gamma}) + \boldsymbol{\gamma}^T\boldsymbol{\Omega}_\lambda(\boldsymbol{\gamma}^0)\boldsymbol{\gamma},$$

where $\boldsymbol{\Omega}_\lambda(\boldsymbol{\gamma}^0) = \frac{1}{2}diag(\frac{p'_\lambda(||\boldsymbol{\gamma}_1^0||)}{||\boldsymbol{\gamma}_1^0||}\mathbf{R}_1, ..., \frac{p'_\lambda(||\boldsymbol{\gamma}_K^0||)}{||\boldsymbol{\gamma}_K^0||}\mathbf{R}_K) = \frac{\lambda}{2}diag(\frac{1}{||\boldsymbol{\gamma}_1^0||}\mathbf{R}_1, ..., \frac{1}{||\boldsymbol{\gamma}_K^0||}\mathbf{R}_K)$. Here we used the REML likelihood to get better estimates for nuisance parameters.

Once the parameter $\boldsymbol{\gamma}$ has been estimated as $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^0$, the REML estimate for σ_0^2 is $\hat{\sigma_0^2} = \frac{1}{n-pK}\mathbf{r}^T\mathbf{V}^{-1}\mathbf{r}$, where $\mathbf{r} = \mathbf{y} - \mathbf{X}_b\hat{\boldsymbol{\gamma}}$. To simplify the computation, we substitute it to the penalized likelihood function to get

$$F(\theta|\boldsymbol{\gamma} = \boldsymbol{\gamma}^0) = log|\mathbf{V}| + log(\mathbf{X}_b^T\mathbf{V}^{-1}\mathbf{X}_b) + (n-pK)log(\mathbf{r}^T\mathbf{V}^{-1}\mathbf{r}) + \boldsymbol{\gamma}^{0T}\boldsymbol{\Omega}_\lambda(\boldsymbol{\gamma}^0)\boldsymbol{\gamma}^0.$$

Thus the REML estimate of θ can be obtain by Newton-Raphson method.

Once the estimates $\hat{\sigma_0^2}$ and $\hat{\theta}$ are obtained, we have

$$F(\boldsymbol{\gamma}|\sigma_0^2 = \hat{\sigma_0^2}, \theta = \hat{\theta}) = \hat{\sigma_0}^{-2}(\mathbf{y} - \mathbf{X}_b\boldsymbol{\gamma})^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}_b\boldsymbol{\gamma}) + \boldsymbol{\gamma}^T\boldsymbol{\Omega}_\lambda(\boldsymbol{\gamma}^0)\boldsymbol{\gamma} + constant$$

We developed an iterative algorithm as following:

Step(1). Initialize $\boldsymbol{\gamma} = \boldsymbol{\gamma}^0$ by simple linear regression.

Step(2). Given $\boldsymbol{\gamma}^{(i)}$, update nuisance parameters θ and σ_0^2 to $\theta^{(i+1)}$ and $\sigma_0^{2(i+1)}$ by the REML estimates as described above.

Step(3). Given $\boldsymbol{\gamma}^{(i)}$, $\theta^{(i+1)}$ and $\sigma_0^{2(i+1)}$, update $\boldsymbol{\gamma}$ to $\boldsymbol{\gamma}^{(i+1)}$ with as

$$\boldsymbol{\gamma}^{(i+1)} = \boldsymbol{\gamma}^{(i)} - (\sigma_0^{-2(i+1)}\mathbf{X}_b^T\mathbf{V}(\theta^{(i+1)})^{-1}\mathbf{X}_b + \boldsymbol{\Omega}_\lambda(\boldsymbol{\gamma}^{(i)}))^{-1}\{-\mathbf{X}_b^T\mathbf{V}(\theta^{(i+1)})^{-1}(\mathbf{y} - \mathbf{X}_b\boldsymbol{\gamma}) + \boldsymbol{\Omega}_\lambda(\boldsymbol{\gamma}^{(i)})\boldsymbol{\gamma}^{(i)}\}.$$

Step(4). Repeat steps (2) and(3) until convergence.

In step (2), if some $||\boldsymbol{\gamma}_j^{(i)}||$ is smaller than a cutoff value ϵ, then we set $\hat{\boldsymbol{\gamma}}_j = \mathbf{0}$. We set ϵ to 10^{-3} in our implementation of the algorithm.

The estimation of parameters $\boldsymbol{\gamma}$, θ and σ_0^2 depends on the tuning parameter λ. How to choose the tuning parameter is crucial for the procedure. The most commonly used methods are the cross-validation and the generalized cross-validation. However, K-fold CV and leave-one-out CV are very intensive in computation. Furthermore, both CV and GCV are created for independent data. Although they can be easily extended to correlated data that are independent between subjects or any higher lever units, the correlation among RIX individuals is not block diagonal. Therefore, it is hard to apply CV or GCV to the data. So we decided to use information criteria to select λ.

The predictor in our model, marker genotypes, are highly correlated due to linkage. Hence an over-fitted model with small λ is more likely to be picked based on any information criteria, AIC or BIC. So we proposed to used a two-stage procedure to select tuning parameters as following:

Step(1). Set a range of values for λ. For any λ, fit the penalized mixed effect model using the algorithm mentioned above, and variable selection is performed since some groups of predictors may shrink to 0.

Step(2). For a given tuning parameter λ, fit a regular mixed effect model using the set of predictors selected in Step(1), for which parameters can be estimated. Then AIC and

BIC are computed for the mixed effect model

$$AIC = -2l + 2q,$$

and

$$BIC = -2l + qlog(n),$$

where l is the log-likelihood, n is the number of subjects and q is the number of predictors in the model.

Step(3). Pick the λ (and the reduced model based on that λ) that minimizes AIC or BIC.

3.3 Simulation Results

We applied the loop design for mating scheme as described by Zou et al. (2005) in simulation studies. We set number of RI lines $L = 100$ and the number of subjects $n = 300$. In each run of simulation, a single chromosome with 51 evenly spaced markers is simulated with 5cM-interval between nearby markers (resulting in a total length of 2.5 Morgan). The marker genotypes are simulated using R/qtl (Broman et al. 2003). We pick the two markers located at 100 cM and 125 cM as two QTL. The functional effect of the two QTL are denoted as β_1 and β_2.

We considered three different functions for the varying coefficient in three scenarios. In each scenario, there are two cases: two QTL are either linked in repulsion or linked in coupling (Broman and Speed 2002), which means that either $\beta_1 = -\beta_2$ or $\beta_1 = \beta_2$, respectively. The functions for varying coefficients in all three scenarios are summarized as below.

Case 1a: $\beta_1(t) = -\beta_2(t) = 1 + 3sin(\frac{\pi t}{30})$.

Case 1b: $\beta_1(t) = \beta_2(t) = 1 + 2sin(\frac{\pi t}{30})$.

Case 2a: $\beta_1(t) = -\beta_2(t) = 1 - 3cos(\frac{\pi t}{60})$.

Case 2b: $\beta_1(t) = \beta_2(t) = 1 - 2cos(\frac{\pi t}{60})$.

Case 3a: $\beta_1(t) = -\beta_2(t) = 1 + \frac{(30-t)^3}{5000}$.

Case 3b: $\beta_1(t) = \beta_2(t) = 1 + \frac{(30-t)^3}{7000}$.

In all scenarios, we set $\mu = 0$, $\sigma_a^2 = 20$, and $\sigma_0^2 = 10$. t is randomly generated from $(0, 60)$. The average heritability for each scenario is between 0.070 and 0.078. For each case, model selection have been performed and parameters have been estimated using the method described above. Two methods, AIC and BIC, are used to select the tuning parameter λ. For comparison, single marker analysis has been performed on each marker. We applied the same method as in the previous chapter to get the empirical threshold value for LOD score, which has been used to access the significance for hypothesis testings in single gene model. Simulations has been performed 100 times for each scenario.

The proportion that each marker was selected into the model has been plotted along the locations of markers in figures 3.1-3.3. The BIC method generally performs better than the AIC method for less false selections, although with a little less correct selections. The penalized likelihood approaches are doing better than single marker analysis in terms of variable selection, especially in cases where effects of two QTL are in liked in repulsion. In cases where effects of two QTL are linked in coupling, they have more accurate QTL mapping positions than the single marker analysis.

The average number of correctly and incorrectly selected variables, and the proportion that the selected model is exactly the true model are recorded in table 3.1. The mean sum of squared distance in the same table is calculated by the average sum of squared distance of any selected marker to the nearest QTL. The AIC method has the highest mean correct selection in cases where effects of two QTL are linked in repulsion, and single marker analysis has highest mean correct selection in cases where effects of two QTL are linked in coupling. In all cases where effects of two QTL are linked in repulsion and one case where effects of two QTL are linked in coupling, the BIC method has the

highest proportion of true model, and the AIC method has the highest proportion of true model in the rest two cases. Overall speaking, the AIC and BIC methods have better performance than the single marker analysis in variable selection, and the BIC method works a little better than the AIC method.

It is probably rare for two functional QTL to be linked in coupling or in repulsion at any t. Therefore, in another set of simulations, we considered four different functions for genetic effects of four QTL. The four QTL are located at 75 cM, 100 cM, 125 cM and 150 cM on a 250 cM chromosome with 50 equally spaced markers, and their effects are β_1, β_2, β_3, β_4, respectively:

$\beta_1(t) = \frac{3}{1+2e^{-0.1t}}$.

$\beta_2(t) = 3sin(\frac{\pi t}{30})$.

$\beta_3(t) = \frac{(30-t)^3}{5000}$.

$\beta_4(t) = arctan(\frac{t-30}{4}) + \frac{\pi}{2}$.

We randomly generated marker genotypes as before with same number of subjects and set $\mu = 0$. We applied two sets of variances $\sigma_a^2 = 20$, $\sigma_0^2 = 40$, and $\sigma_a^2 = 10$, $\sigma_0^2 = 20$, where the average heritability is 0.10 or 0.18, respectively. Model selection have been performed by penalized likelihood method with the two approaches, AIC and BIC, as well as single marker analysis. Simulations has been performed 100 times. The proportion that each marker was selected into the model has been plotted along the locations of markers in figure 3.4.

To evaluate the performance of the variable selection methods, we compared the number of true discoveries and false discoveries across different size of the tuning parameter λ for the penalized likelihood method, and across different cutoffs of LOD scores for single marker analysis. We obtain final models for each method using a series of cutoffs. We determine the number of true discoveries in the final model as follows. For each true QTL, we check whether there is any marker in the final model falls in a window of

certain size. We applied three different window sizes, 0 cM (no window), 10 cM and 20 cM. For example, if the window size is 10 cM, then any any marker in the final model located within 5 cM to the QTL is considered as a true discovery. If there is no such marker, there is no true discovery for this QTL. When the window size is not 0 cM, if there is at least one such marker, the nearest one to the QTL is recorded as a true discovery and is excluded from the true discovery when searching of other QTL. After the true discoveries of all the QTL are identified, the remaining markers in the final model are defined as false discoveries, which can be further divided into two categories: false discoveries linked to at least one QTL (linked false discoveries) and false discoveries unlinked to any QTL (unlinked false discoveries). A linked false discovery is a marker within the genetic distance (5 cM or 10 cM) to one QTL and is not counted as true discovery.

We summarized the results of each method by an ROC-like curve (Sun et al. 2010) that plots the mean number of true discoveries versus the mean number of false discoveries across a serial of cutoff values in figures 3.5 and 3.6. The method with ROC-like curves closer to the upper-left corner of the plot has better variable selection performance because it have more true discoveries, given the same number of false discoveries. If the linked false discoveries are counted as false discoveries, the performance of the penalized likelihood method performs substantially better than the single marker analysis for any window size. If the linked false discoveries are excluded from false discoveries, then the penalized likelihood method outperforms the single marker analysis in ROC-like curves when the window size is small (10 cM), but the performances of the two methods are not well-separated when the window size is big (20 cM). As expected, ROC-like curves can be improved with higher heritability (lower variances). Overall speaking, the multiple-loci mapping by penalized likelihood performs better than single maker analysis in QTL identification.

3.4 Discussion

In this chapter, we extended the single QTL mapping for functional traits to multiple QTL mapping through a penalized likelihood approach by group selection of coefficients associated with genes. We proposed a two-stage selection procedure to reduce the impact of high correlation among marker genotypes on variable selection. Simulation studies showed that the multiple QTL mapping method generally performs better than single QTL model. Given the same false discovery, the multiple QTL model identified more true QTL than single marker analysis.

Since the RIX data has correlated residuals, we used a mixed effect model and applied penalty function to the likelihood of the mixed effect model. We estimated the nuisance parameters by REML, which is not applicable to high dimension data, where the number of markers more than the number of subjects. To extend this method to high dimension data,a penalized weighted least squares approach (Wang et al. 2008) might be considered, where a weight matrix is incorporated similar to a working correlation in GEE. It is possible to extend the method to include epistasis into the model, although epistasis for functional traits could be complicated.

One problem with the multiple QTL mapping by group selection is that when the penalty is small, some markers that are far away from QTL has a better chance to be selected than in single marker analysis. This implies that choosing appropriate tuning parameter might be critical in variable selection for multiple-loci mapping.

	Case 1a	Case 1b	Case 2a	Case 2b	Case 3a	Case 3b
Mean Correct selection						
Single-marker	1.38	1.87	1.41	1.85	1.28	1.86
Multiple-AIC	1.97	1.76	1.99	1.69	1.95	1.71
Multiple-BIC	1.68	1.08	1.78	1.21	1.58	1.15
Mean Incorrect selection						
Single-marker	0.95	4.32	1	4.54	1.29	4.78
Multiple-AIC	2.12	2.25	2.12	1.98	2.41	1.68
Multiple-BIC	0.29	0.34	0.24	0.28	0.33	0.27
Proportion of true model						
Single-marker	0.16	0.04	0.13	0.03	0.07	0.06
Multiple-AIC	0.31	0.23	0.35	0.27	0.27	0.23
Multiple-BIC	0.57	0.17	0.66	0.23	0.52	0.26

Table 3.1: Performance of model selection for the three methods

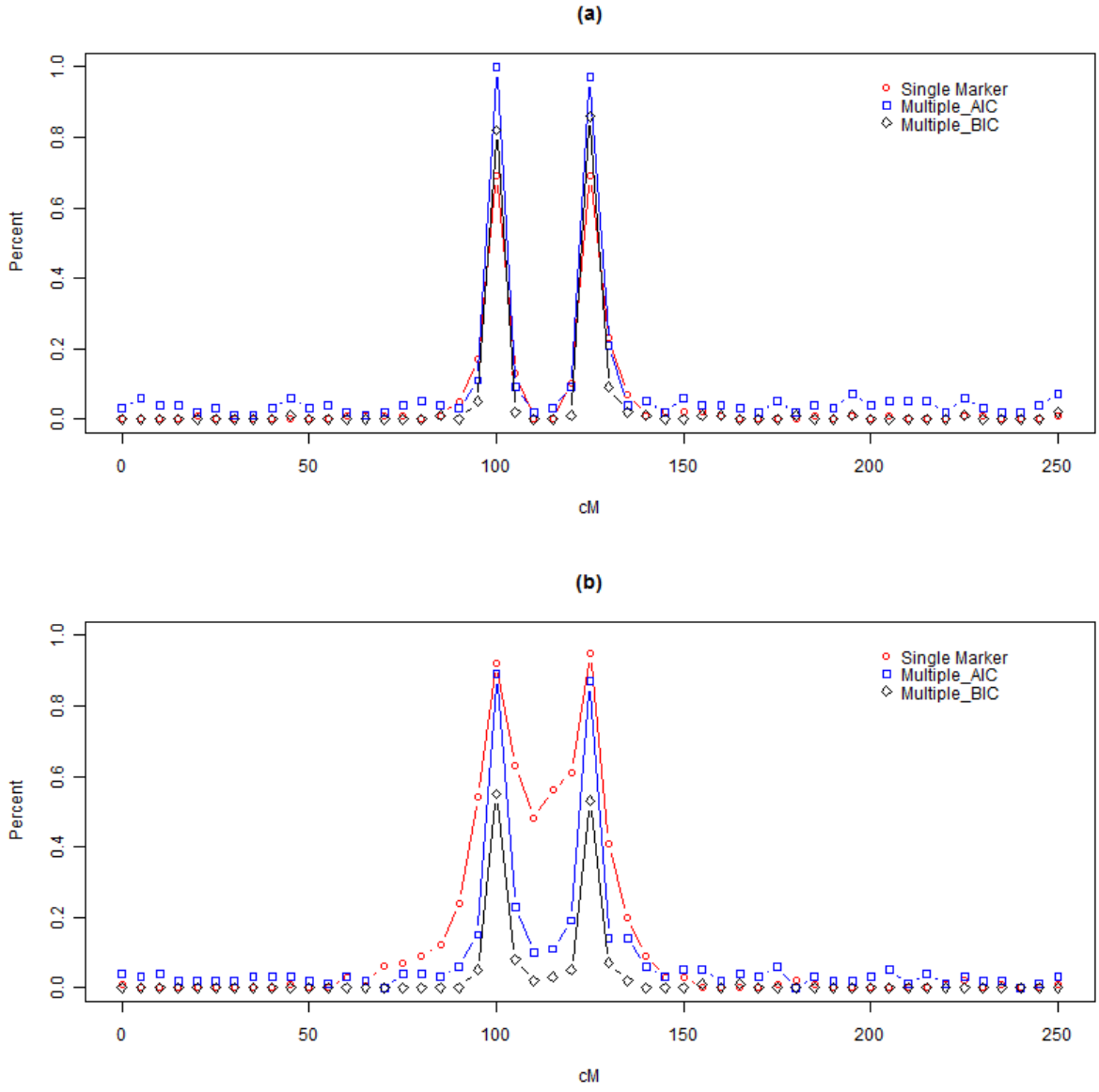

Figure 3.1: Proportion of selection for (a) Case 1a and (b) Case 1b

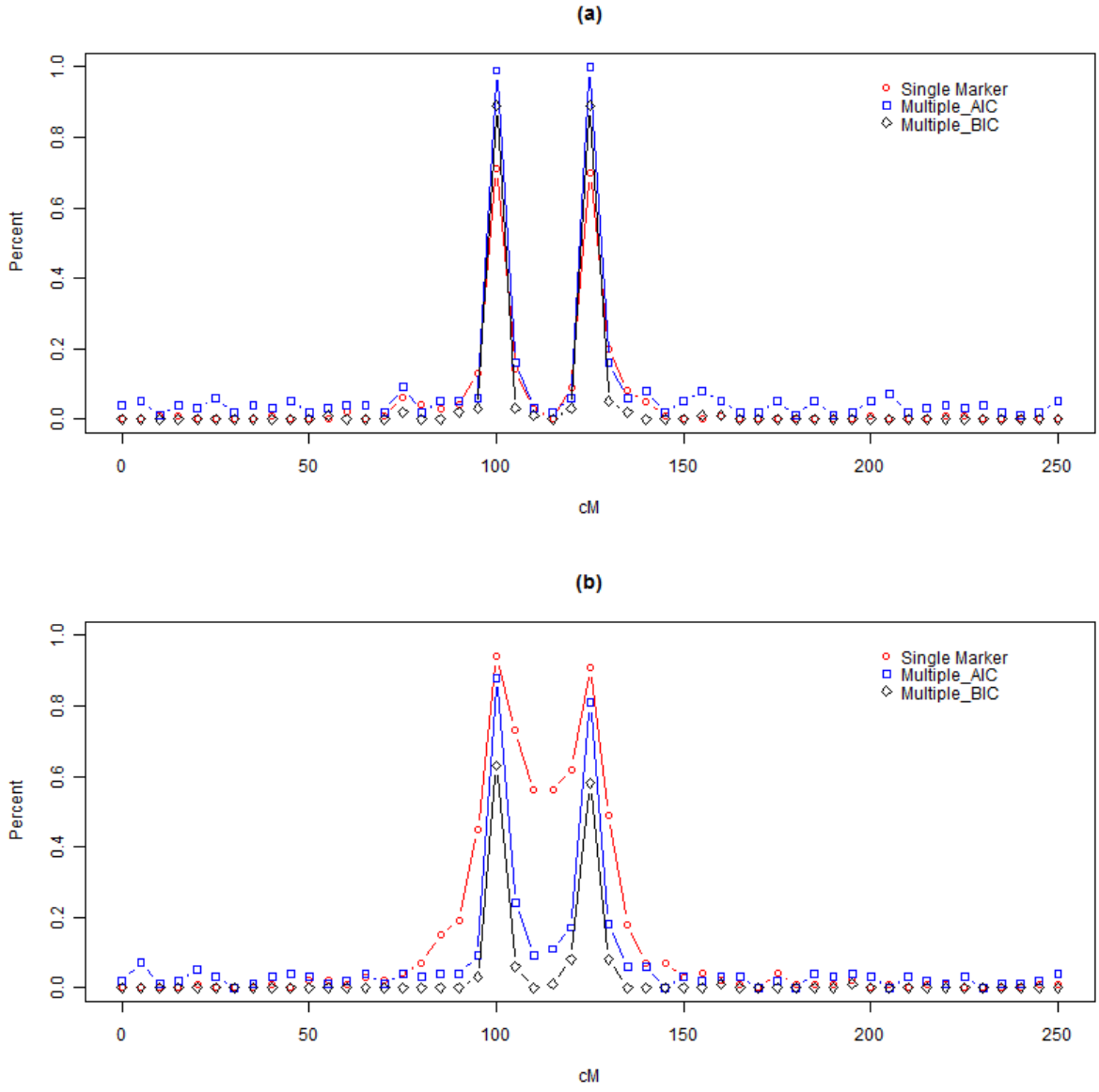

Figure 3.2: Proportion of selection for (a) Case 2a and (b) Case 2b

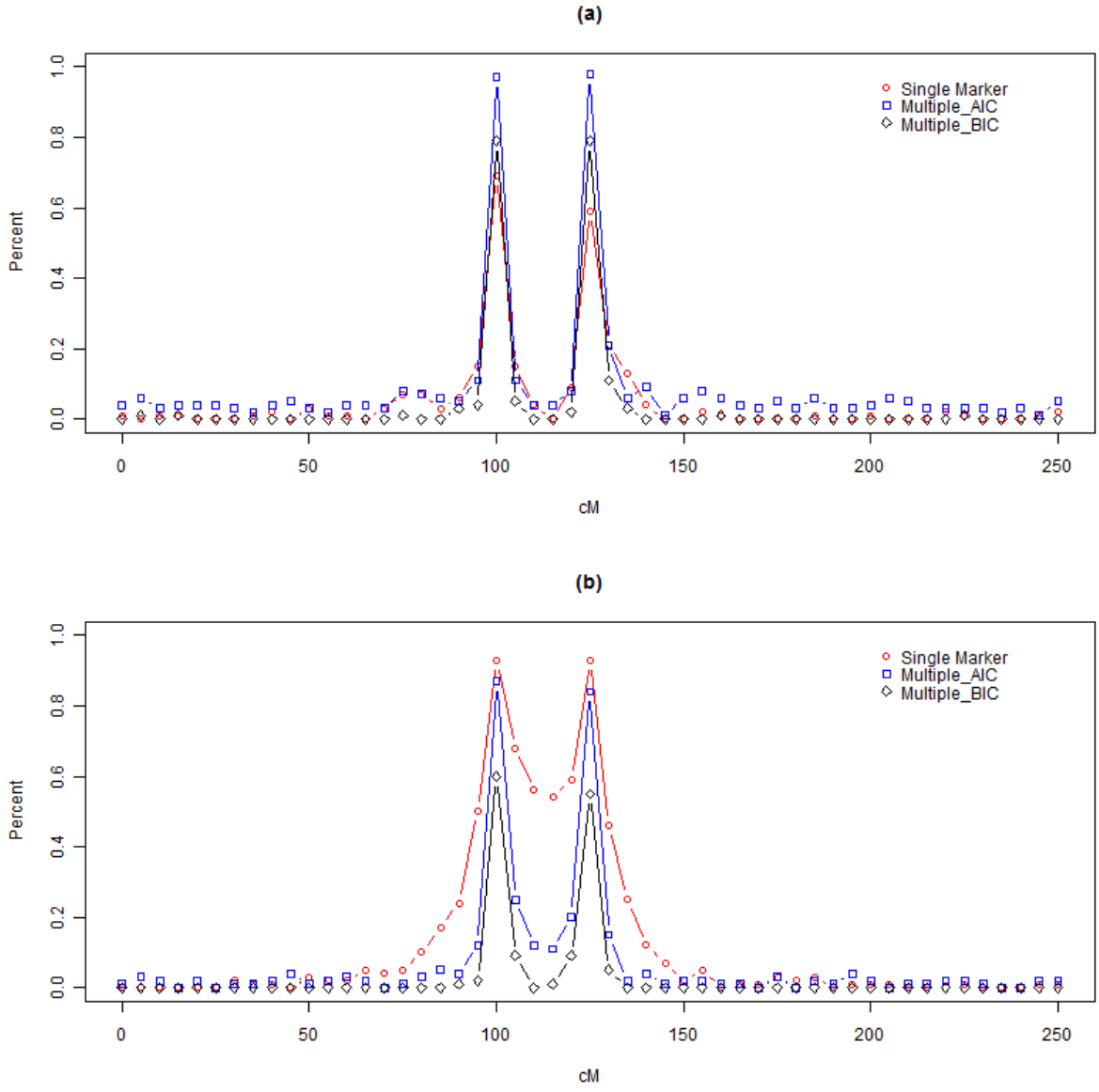

Figure 3.3: Proportion of selection for (a) Case 3a and (b) Case 3b

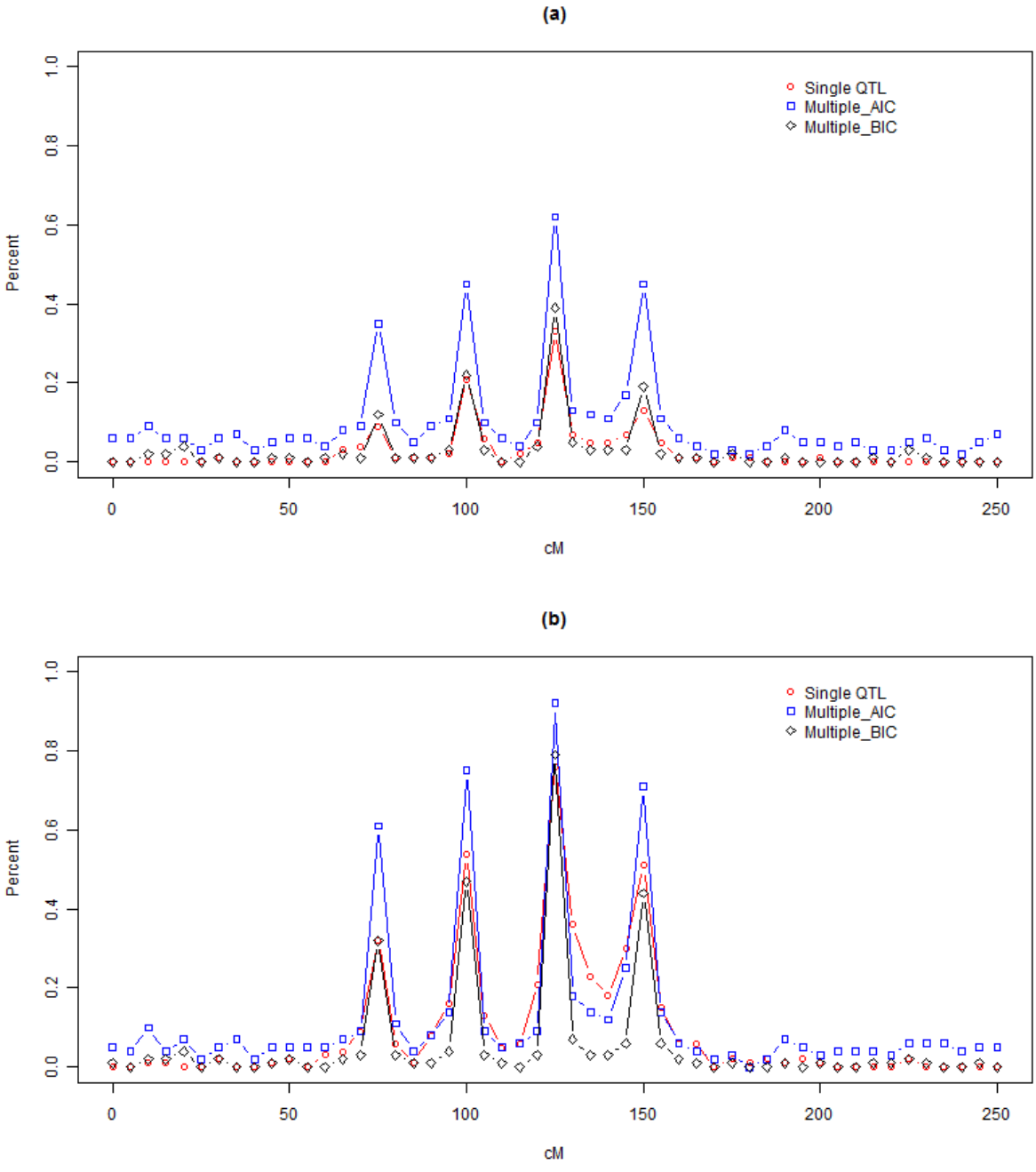

Figure 3.4: Proportion of selection for the three methods

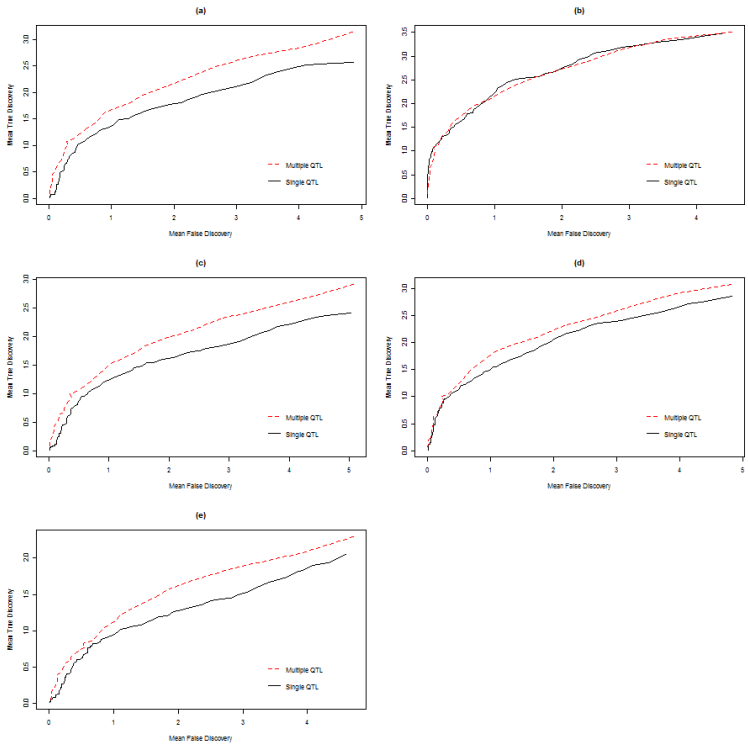

Figure 3.5: ROC-like plots with $\sigma_a^2 = 20$ and $\sigma_0^2 = 40$

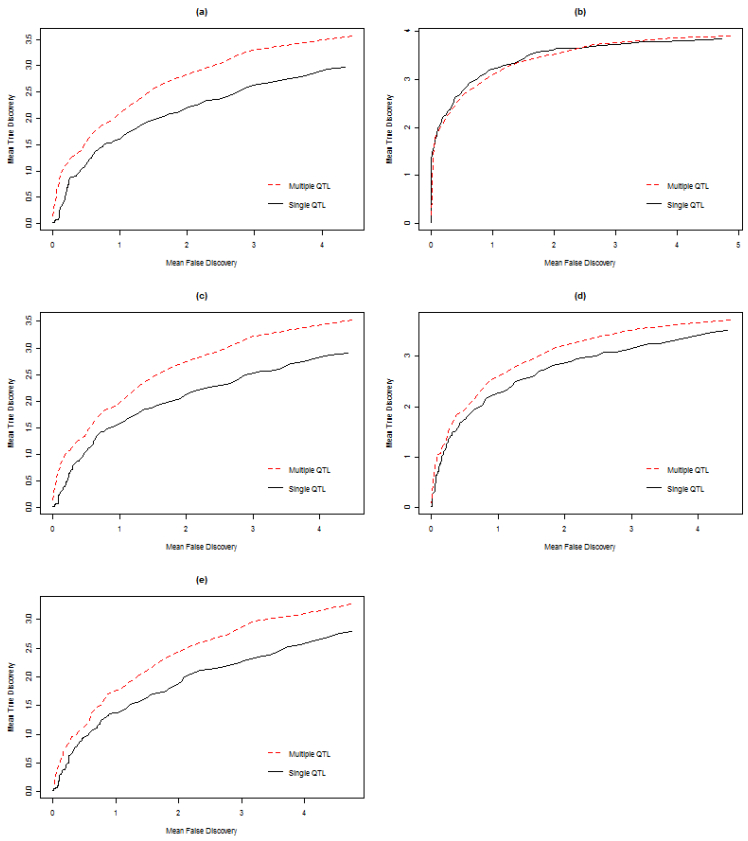

Figure 3.6: ROC-like plots with $\sigma_a^2 = 10$ and $\sigma_0^2 = 20$

Chapter 4

Mapping Multiple QTL for Longitudinal Traits by Model Selection

4.1 Introduction

The functional traits studies in the previous chapters can be generalized to traits with repeated measures on the same individual over time. Many important traits, such as body size or daily milk yield, are expressed continuously throughout life or for a period of life. These traits, called longitudinal traits, are traditionally analyzed in terms of a set of heritabilities at each age and correlations between different ages, but with no consideration of the time dependent continuity that must exist between successive ages (Zhao et al. 2005). Wu and colleagues (Ma et al. 2002, Wu et al. 2002, Wu et al. 2004, Zhao et al. 2005, Lin and Wu 2006) developed the functional mapping approach, which provided a useful framework for genetic mapping through mean and covariance modeling of longitudinal traits.

The mean parameters can be modeled by some parametric function such as sigmoidal or logistic function. For example, Ma et al. (2002) used diameter growth of poplar as

an example of functional trait and modeled genetic effects by a logistic function. A more flexible way to model functional genetic effects nonparametrically by functional approximation. Various basis systems can be used, and the most common choice is the B-spline bases (He and Shi 1998, Pittman 2002, Huang et al. 2004, Wang et al. 2008, Yang wt. al 2009).

Another important issue in longitudinal analysis is how to model the variance structure. The autoregressive models are a class of covariance models widely used in longitudinal data modeling (Diggle et al. 2002). A first-order stationary autoregressive model, or AR(1) model, has a simple structure, with only 2 parameters. Its inverse and determinant have closed forms, which makes computation easier and faster. However, an AR(1) model assumes the longitudinal data has stationary variance and covariance, which is questionable in a lot of cases (Zhao et al. 2005). To deal with the heteroscedastic problem of the residual variance, one approach is to model the residual variance by a parametric function of time (Pletcher and Geyer 1999). But this approach needs to implement additional parameters for characterizing the age-dependent change of the variance. Another approach is to use transform the data by the transform-both-sides method (Wu et al. 2004) and then use the AR(1) model on the transformed data to achieve stationary variance. However, the stationary covariance assumption is still a problem.

Antedependence models are useful generalizations stationary autoregressive models that are able to model both nonstationary variance and correlation functions. The antedependence model was originally proposed by Gabriel (1962), which assumes serial correlation within subjects like the autoregressive model but allows for nonstationary variation. It states that an observation at a particular time t depends on the previous ones, with the degree of dependence decaying with time lag. If an observation at time t is independent of all observations before $t - r$, this antedependent model is called rth-order, or AD(r). A T-variate normal random vector $\mathbf{y} = (y_1, ..., y_T)^T$ with mean

$\boldsymbol{\mu} = (\mu_1, ..., \mu_T)^T$ follows an AD(r) model if

$$\begin{array}{l} y_1 = \mu_1 + \epsilon_1 \\ y_t = \mu_t + \sum_{k=1}^{r^*} \phi_{kt}(y_{t-k} - \mu_{t-k}) + \epsilon_t \quad t = 2, 3, ..., T \end{array} , \tag{4.1}$$

where $r^* = min(r, t-1)$, the ϵ_t's are independent random variables following $N(0, \sigma_t^2)$, and ϕ_{kt}'s are such that the covariance matrix, Σ, is positive definite. It is easy to observe that both variance and correlation are nonconstant over time, as long as some $\phi_{kt} \neq 0$. Antedependence models are very useful for longitudinal data exhibiting heterogeneous variances and nonstationary serial correlation, such as data in growth studies (Nunez-Anton and Zimmerman, 2000). However, the covariance of an unstructured AD(r) model, UAD(r), is specified by $(r+1)(2T-r)/2$ parameters, which is not so parsimonious.

To make the antedependence model more parsimonious and useful, Nunez-Anton (1997) and Nunez-Anton and Zimmerman (2000) proposed structured antedependence (SAD) models, which incorporate some structural forms of nonstationary into AD models. Denoting the measurement times of any subject as $t_1 < t_2 < ... < t_T$, an rth-order SAD, or SAD(r), model can be specified as

$$\begin{array}{l} \phi_{i-k,i} = f(t_i, t_{i-k}; \lambda_k) \\ \sigma_{ii} = \sigma^2 g(t_i; \boldsymbol{\psi}) \end{array} ,$$

where $\sigma^2 > 0$, $\boldsymbol{\psi}$, $\lambda_1, ..., \lambda_r$ are parameters such that the covariance matrix Σ is positive definite, $f(.)$ and $g(.)$ are specified functions. Regardless of the forms for $f(.)$ and $g(.)$, the SAD(r) model above only involves as less as $r+2$ parameters. Therefore, the SAD models can be much more parsimonious than the UAD models when T is not too small. Another good property of the SAD model is that its inverse has a simple form and is very easy to calculate.

The remainder of the chapter is organized as follows. In the method section, we extended the penalized likelihood method for multiple functional QTL mapping to longitudinal data with nonstationary covariance. In the results section, we presented simulation results for the performance of the model selection method. In the discuss section, we conclude and discuss the implications of our model.

4.2 Methods

4.2.1 Model

We are interested in longitudinal traits influenced by genetic effects changing over time, which we model as

$$y_{il} = \mu(t_{il}) + \sum_{j=1}^{p} x_{ij}\beta_j(t_{il}) + \epsilon_{il},$$

where t_{il} is the time of the lth measurement for the ith individual; y_{il} is the measure of the phenotype of individual i at time t_{il}; x_{ij} is the genotype of the ith individual at the jth marker, coded as -1, 0 or 1 for genotypes aa, Aa and AA, respectively; ϵ_{il} is the random error with $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ...\epsilon_{im})^T$ following $N(\mathbf{0}, \boldsymbol{\Sigma}_i)$; t_{il} affects the coefficients of predictors x_{ij} through any function $\beta_j()$, for which we model nonparametrically using basis expansion. For simplicity, we assume every subject has same number of measurement m, and the total number of observations is $N = nm$.

The coefficient $\beta_j(t_l)$ can be approximated by

$$\beta_j(t_{il}) \approx \sum_{k=1}^{K} \gamma_{jk} B_{jk}(t_{il}),$$

where $B_{jk}(t)$'s are B-spline basis functions. The smoothness of the coefficient functions modeled by B-splines are controlled by the parameter $K = n_j + d + 1$, where n_j is the number of interior knots and d is the degree of spline. The interior knots of the splines can be either equally spaced or placed on the sample quantiles of the data, so that there are

about the same number of observations between any two adjacent knots.We use equally spaced knots for all numerical examples in this article. Thus $B_{jk}(t)$ is predetermined for any given t.

We can rewrite the model above into matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_n)^T$ with $\mathbf{y}_i = (y_{i1}, ..., y_{im})^T$; the intercept $\mu(t_{il}) = \sum_{k=1}^{K} \gamma_{0k}(t_{il})B_{0k}(t_{il})$; $\boldsymbol{\gamma} = (\gamma_{01}...\gamma_{0K}, ..., \gamma_{p1}...\gamma_{pK})^T$ is a vector of $(p+1)K$ parameters; and $\mathbf{X} = (\mathbf{X}_1^T, ..., \mathbf{X}_n^T)^T$ is the corresponding $N \times ((p+1)K)$ design matrix for the time varying effects, where $\mathbf{X}_i$ has the form

$$\mathbf{X}_i = \begin{pmatrix} B_{01}(t_{i1}) & \dots & B_{0K}(t_{i1}) & \dots & x_{ip}B_{p1}(t_{i1}) & \dots & x_{ip}B_{pK}(t_{i1}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B_{01}(t_{im}) & \dots & B_{0K}(t_{im}) & \dots & x_{ip}B_{p1}(t_{im}) & \dots & x_{ip}B_{pK}(t_{im}) \end{pmatrix}.$$

The random error $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, ..., \boldsymbol{\epsilon}_n^T)^T$ can be described by the SAD(r) model. For the SAD(r) model (4.1), Nunez-Anton and Zimmerman (2000) suggested a typical choice of $f(.)$ can be an exponential function as

$$\phi_{i-k,i} = f(t_i, t_{i-k}; \lambda_k) = exp\{-\lambda_k(t_i - t_{i-k})\}.$$

Nunez-Anton and Zimmerman (2000) also suggested to model the $log\sigma^2(t_i)$ by some polynomial functions of t_i, and $g(t_i; \boldsymbol{\psi})$ becomes an exponential function. We simply use a quadratic function to model $log\sigma^2(t_i)$, so that $\sigma^2(t_i)$ can be expressed as

$$\sigma^2(t_i) = exp\{a + bt_i\} = \sigma^2 exp\{bt_i\},$$

where $\sigma^2 = e^a$, a and b are unknown parameters. With such choices of $f(.)$ and $g(.)$,

only $r + 2$ parameters are involved in the model.

For simplicity, we choose $r = 1$, and hence the residual covariance matrix $\mathbf{\Sigma}$, with an SAD(1) model, can be expressed as

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{G}\mathbf{A}^T,$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \phi_{1,2} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \phi_{1,T} & \phi_{2,T} & \dots & 1 \end{bmatrix},$$

and

$$\mathbf{G} = \begin{bmatrix} \sigma^2(t_1) & 0 & \dots & 0 \\ 0 & \sigma^2(t_2) & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma^2(t_T) \end{bmatrix}.$$

The inverse of the matrix $\mathbf{A}$ is

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ -\phi_{1,2} & 1 & 0 & \dots & 0 & 0 \\ 0 & -\phi_{2,3} & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & -\phi_{T-1,T} & 1 \end{bmatrix}.$$

Hence it is very easy to compute the inverse for the covariance matrix $\mathbf{\Sigma}$, as $\mathbf{\Sigma}^{-1} = \mathbf{A}^{-1^T}\mathbf{G}^{-1}\mathbf{A}^{-1}$.

The SAD(1) model can be further simplified, if times of repeated measurements are equally spaced and innovation variances are constant over time points as introduced by

Jaffrezic et al. (2003), and the measurements can be expressed as

$$\begin{aligned} y_1 &= \mu_1 + \epsilon_1 \\ y_t &= \mu_t + \phi(y_{t-1} - \mu_{t-1}) + \epsilon_t \quad t = 2, 3, ..., T \end{aligned},$$

where ϵ_t follows $N(0, \sigma^2)$ with constant innovation variance σ^2. This simple SAD(1) model only involves two parameters σ^2 and ϕ, and hence it is very parsimonious. The analytical forms for variance and covariance functions of this model can be derived as

$$\begin{aligned} \sigma_{ii} &= \tfrac{1-\phi^{2i}}{1-\phi^2}\sigma^2 \\ \sigma_{i-k,i} &= \phi^k \tfrac{1-\phi^{2(i-k)}}{1-\phi^2}\sigma^2 \end{aligned}.$$

It can be easily seen that both variance and correlation functions are non-stationary for the SAD(1) model, even with constant innovation variance σ^2 and constant antedependent coefficient ϕ.

We can write the covariance for the residual $\boldsymbol{\epsilon}_i$ as a function of nuisance parameters, $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{V}_i = \sigma^2 \mathbf{A}(\phi)\mathbf{G}(b)\mathbf{A}(\phi)^T$, where σ^2, ϕ and b are parameters for covariance and the matrices $\mathbf{A}$ and $\mathbf{G}$ have the form as displayed above. We observe that $\mathbf{y}_i$ follows $N(\mathbf{X}_i\boldsymbol{\gamma}, \boldsymbol{\Sigma}_i)$. so the restricted/residual maximum likelihood (REML) can be expressed as

$$-2l_R(\boldsymbol{\gamma}, \sigma^2, \phi, b) = \sum_{i=1}^{n} log|\mathbf{V}_i| + (N - pK)log(\sigma^2) + \sum_{i=1}^{n} |\mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i| + \sigma^{-2} \sum_{i=1}^{n} \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i,$$

where $\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\gamma}$.

For variable selection, we can impose penalty functions, $p_\lambda(||\boldsymbol{\gamma}_j||) = \lambda||\boldsymbol{\gamma}_j||$ to the log-likelihood function above. Then the objective function can be expressed as

$$F(\boldsymbol{\gamma}, \sigma^2, \phi, b) = \sum_{i=1}^{n} log|\mathbf{V}_i| + (N - pK)log(\sigma^2) + \sum_{i=1}^{n} |\mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i| + \sigma^{-2} \sum_{i=1}^{n} \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i + \sum_{j=1}^{p} p_\lambda(||\boldsymbol{\gamma}_j||),$$

where $\boldsymbol{\gamma}_j = (\gamma_{j1}, ..., \gamma_{jK})^T$ and $||\boldsymbol{\gamma}_j||$ denotes some L_2-norm of $\boldsymbol{\gamma}_j$. As proposed by Wang et al. (2008), $||\boldsymbol{\gamma}_j||$ is defined as the L_2-norm of the function $\beta_j(t) = \sum_{k=1}^{K} \gamma_{jk} B_{jk}(t)$. The squared L_2-norm can be written as quadratic form $||\boldsymbol{\gamma}_j||^2 = \boldsymbol{\gamma}_j^T \mathbf{R}_j \boldsymbol{\gamma}_j$, where $\mathbf{R}_j = (r_{uv})_{K\times K}$ is a matrix with entries $r_{uv} = \int_{t_1}^{t_2} B_{ju}(t) B_{jv}(t) dt$ with (t_1, t_2) being the range of t.

4.2.2 Computational Algorithm

For the likelihood based approach, we applied the quadratic approximations method by Fan and Li (2001), and the objective function can be approximately expressed as

$$F(\boldsymbol{\gamma}, \sigma^2, \phi, b) = \sum_{i=1}^{n} log|\mathbf{V}_i| + (N-pK) log(\sigma^2) + \sum_{i=1}^{n} |\mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i| + \sigma^{-2} \sum_{i=1}^{n} \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i + \boldsymbol{\gamma}^T \boldsymbol{\Omega}_\lambda(\boldsymbol{\gamma}^0) \boldsymbol{\gamma},$$

where $\boldsymbol{\Omega}_\lambda(\boldsymbol{\gamma}^0) = \frac{1}{2} diag(\frac{p'_\lambda(||\boldsymbol{\gamma}_1^0||_2)}{||\boldsymbol{\gamma}_1^0||_2} \mathbf{R}_1, ..., \frac{p'_\lambda(||\boldsymbol{\gamma}_K^0||_2)}{||\boldsymbol{\gamma}_K^0||_2} \mathbf{R}_K) = \frac{\lambda}{2} diag(\frac{1}{||\boldsymbol{\gamma}_1^0||_2} \mathbf{R}_1, ..., \frac{1}{||\boldsymbol{\gamma}_K^0||_2} \mathbf{R}_K)$.

Once the parameter $\boldsymbol{\gamma}$ has been estimated as $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^0$, the REML estimate for σ^2 is $\hat{\sigma^2} = \frac{1}{n-pK} \sum_{i=1}^{n} \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i$, where $\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i \boldsymbol{\gamma}^0$. To simplify the computation, we substitute it into the penalized likelihood function to get

$$F(\phi, b | \boldsymbol{\gamma} = \boldsymbol{\gamma}^0) = \sum_{i=1}^{n} log|\mathbf{V}_i| + log(\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i) + (n-pK) log(\sum_{i=1}^{n} \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i) + \boldsymbol{\gamma}^{0T} \boldsymbol{\Omega}_\lambda(\boldsymbol{\gamma}^0) \boldsymbol{\gamma}^0.$$

Thus the REML estimate of nuisance parameters ϕ and b can be obtain by Newton-Raphson method.

Once the estimates $\hat{\sigma^2}$, $\hat{\phi}$ and $\hat{b}$ are obtained, we have

$$F(\boldsymbol{\gamma} | \sigma^2 = \hat{\sigma^2}, \phi = \hat{\phi}, b = \hat{b}) = \hat{\sigma}^{-2} (\mathbf{y} - \mathbf{X}\boldsymbol{\gamma})^T \hat{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\gamma}) + \boldsymbol{\gamma}^T \boldsymbol{\Omega}_\lambda(\boldsymbol{\gamma}^0) \boldsymbol{\gamma} + constant,$$

where $\mathbf{V} = Blockdiag(\mathbf{V}_1, ..., \mathbf{V}_N)$.

We can use iterative algorithm as following:

Step(1). Initialize $\boldsymbol{\gamma} = \boldsymbol{\gamma}^0$ by simple linear regression.

Step(2). Given $\boldsymbol{\gamma}^{(i)}$, update nuisance parameters ϕ, b and σ^2 to $\phi^{(i+1)}$, $b^{(i+1)}$ and $\sigma^{2(i+1)}$ by the REML estimates as described above.

Step(3). Given $\boldsymbol{\gamma}^{(i)}$, $\phi^{(i+1)}$, $b^{(i+1)}$ and $\sigma^{2(i+1)}$, update $\boldsymbol{\gamma}$ to $\boldsymbol{\gamma}^{(i+1)}$ with as

$$\boldsymbol{\gamma}^{(i+1)} = \boldsymbol{\gamma}^{(i)} - \{\sigma^{-2(i+1)}\mathbf{X}^T\mathbf{V}(\phi^{(i+1)}, b^{(i+1)})^{-1}\mathbf{X} + \boldsymbol{\Omega}_\lambda(\boldsymbol{\gamma}^{(i)})\}^{-1}$$
$$\{-\sigma^{-2(i+1)}\mathbf{X}^T\mathbf{V}(\phi^{(i+1)}, b^{(i+1)})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\gamma}) + \boldsymbol{\Omega}_\lambda(\boldsymbol{\gamma}^{(i)})\boldsymbol{\gamma}^{(i)}\}.$$

Step(4). Repeat steps (2) and(3) until convergence.

The predictor in our model, marker genotypes, are highly correlated due to linkage. Hence an over-fitted model with small λ is more likely to be picked based on any information criteria, AIC or BIC. So we proposed to used a two-stage procedure to select tuning parameters as following:

Step(1). Set a range of values for λ. For any λ, fit the penalized mixed effect model using the algorithm mentioned above, and variable selection is performed since some groups of predictors may shrink to 0.

Step(2). For a give tuning parameter λ, fit a regular mixed effect model using the set of predictors selected in Step(1), for which parameters can be estimated. Then AIC and BIC are computed for the mixed effect model

$$AIC = -2l + 2q,$$

and

$$BIC = -2l + qlog(n),$$

where l is the log-likelihood, n is the number of subjects and q is the number of predictors.

Step(3). Pick the λ (and the reduced model based on that λ) that minimizes AIC or BIC.

4.3 Simulation Results

In simulation studies, a single chromosome with 51 evenly spaced markers is simulated with 5cM-interval between nearby markers (resulting in a total length of 2.5 Morgan). There are two QTL in the chromosome, located at 100 cM and 125 cM. The effect of the two QTL are denoted as β_1 and β_2. 300 progeny from F_2 intercross are created with their marker genotypes generated by R/qtl (Broman et al. 2003). We considered two different functions for the varying coefficient. The set of times $\mathbf{t}_i = (t_{i1}, .., t_{im})$ is randomly generated from $(0, 60)$, with $m = 5$. The functions for varying coefficients are summarized as below.

Case 1: $\beta_1(t) = -\beta_2(t) = 1 + 3sin(\frac{\pi t}{30})$.

Case 2: $\beta_1(t) = -\beta_2(t) = 1 + \frac{(30-t)^3}{5000}$.

We assume the covariance matrix is SAD(1) with $\sigma^2 = 30$ and $\phi = 0.7$. We applied the penalized likelihood model and modeled the covariance with four different structures: AR(1), compound symmetry, independent and SAD(1). Single marker analysis has been performed on each marker using the generalized least squares method with the correct variance structure, and the significance levels for hypothesis testing are corrected by Bonferroni method. Simulations has been performed for 100 runs.

The proportion that each marker was selected into the model, for the three methods, has been plotted along the locations of markers in figure 4.1. The AIC method selects substantially more correct variables than the method BIC or single marker analysis, and BIC method eliminates more incorrect variables than the AIC method or single marker analysis. The proportion that each marker was selected into the model, for the different covariance models, has been plotted along the locations of markers in figures 4.2-4.3. When the covariance models is SAD(1), the true covariance model, with the AIC method, has better performance than other covariance models for less incorrect selections. However, there is no such difference in variable selection with the BIC method

for all covariance models.

To evaluate the performance of the variable selection methods, we performed another set of simulations, where we created four QTL located at 50 cM, 75 cM, 150 cM and 175 cM on a 250 cM chromosome with 50 equally spaced markers. The genetic effects β_1, β_2, β_3 and β_4 are created same as in chapter 2:

$\beta_1(t) = \frac{3}{1+2e^{-0.1t}}$.

$\beta_2(t) = 3sin(\frac{\pi t}{30})$.

$\beta_3(t) = \frac{(30-t)^3}{5000}$.

$\beta_4(t) = arctan(\frac{t-30}{4}) + \frac{\pi}{2}$.

300 progeny from F_2 intercross are simulated with their marker genotypes generated by R/qtl. We assume the covariance matrix is SAD(1) with $\phi = 0.7$ and $\mu(t) = 0$. We run the simulations for two sets of variances $\sigma^2 = 120$ and $\sigma^2 = 80$, with the average heritability is 0.08 or 0.10, respectively. Model selection have been performed with the penalized likelihood method with the two approaches, AIC and BIC, as well as single marker analysis. Simulations has been performed 100 times for each scenario.

To evaluate the performance of the variable selection methods, we compared the number of true discoveries and false discoveries across different size of the tuning parameter λ for the penalized likelihood method, and across different cutoffs of LOD scores for single marker analysis. We applied variable selection and single marker analysis, and get final models using a series of cutoffs for a set of window sizes (0 cM, 10 cM and 20 cM). We determine the number of true discoveries in the final model as described in the previous chapter.

The results of each method are summarized by an ROC-like curve that plots the mean number of true discoveries versus the mean number of false discoveries across a series of cutoff values in figures 4.4 and 4.5. The penalized likelihood method outperforms the single marker analysis for any window size in the ROC-like curves, no matter the linked false discoveries are counted as false discoveries or not. Therefore, the multiple-loci

mapping by penalized likelihood works better than single maker analysis in detecting QTL.

The performance of variables selection can be affected if the covariance model is mis-specified. To study such effects of covariance models, we conducted a set of simulations. The setting of simulation studies is similar to cases 1 and 2, with the same number of subject, and locations of markers and QTL. The set of times $\mathbf{t}_i = (t_{i1}, .., t_{im})$ is simulated in the same way as cases 1 and 2. The functions for varying coefficients are summarized as below.

Case 3: $\beta_1(t) = -\beta_2(t) = 1 + 3sin(\frac{\pi t}{30})$.

Case 4: $\beta_1(t) = -\beta_2(t) = \frac{3}{1+e^{-0.1t}}$.

We assume covariance matrices are SAD(1) model with the parameters: $\sigma^2 = 15$ and $\phi = 0.7$. We considered three different values for parameter b:

Case a: $b = 0$; the innovation variance is constant over time.

Case b: $b = 0.01$.

Case c: $b = 0.02$.

100 runs of simulations have been performed for each case.

Two covariance models were compared, Model 1: innovation variance is constant over time, and Model 2: log of innovation variance is a linear function of time. The models are compared by information criteria, AIC and BIC. Table 4.1 shows the proportion of simulation runs where the covariance model 2 is preferred over the covariance model 1 based on any one of the two information criteria, in all simulation cases. In cases where $b = 0$, both information criteria prefers the covariance model 1. Both information criteria tends to prefer the covariance model 2 as $|b|$ increases. Figure 4.6 shows that when $b = 0$, model 1 is the same as, or a little better than, model 2 in terms of model selection. When $b \neq 0$, the model 2 is better than model. The advantage of the model 2 is more obvious with the increase of $|b|$.

4.4 Discussion

We extended the multiple mapping by model selection into longitudinal traits with repeated measurements, and applied a nonstationary model, the antedependence model, for the covariance structure. We showed, by simulation studies, that multiple QTL mapping by penalized likelihood generally performs better than single QTL model for longitudinal data. The efficiency of multiple QTL mapping will be reduced if the covariance structure is misspecified. We also showed that information criteria can be used to pick the appropriate covariance model among candidate models.

When the correlations between subjects have no obvious form, the penalized GEE approach (Fu 2003) can be used to perform variable selection for multiple QTL mapping with repeated measures.

Recently, Huang et. al. (2006) proposed covariance matrix selection and estimation via penalized normal likelihood, developed from the modified Cholesky decomposition advocated by Pourahmadi (1999, 2000). Kou and Pan (200) introduced variable selection for joint mean and covariance models via penalized likelihood. It will be interesting to employ it to multiple QTL mapping for longitudinal traits.

	Criterion	$b = 0$	$b = 0.01$	$b = 0.02$
Case5	AIC	0.16	0.92	1
	BIC	0.11	0.92	1
Case6	AIC	0.17	0.95	1
	BIC	0.12	0.92	1

Table 4.1: Proportion of simulation runs that prefer model 2

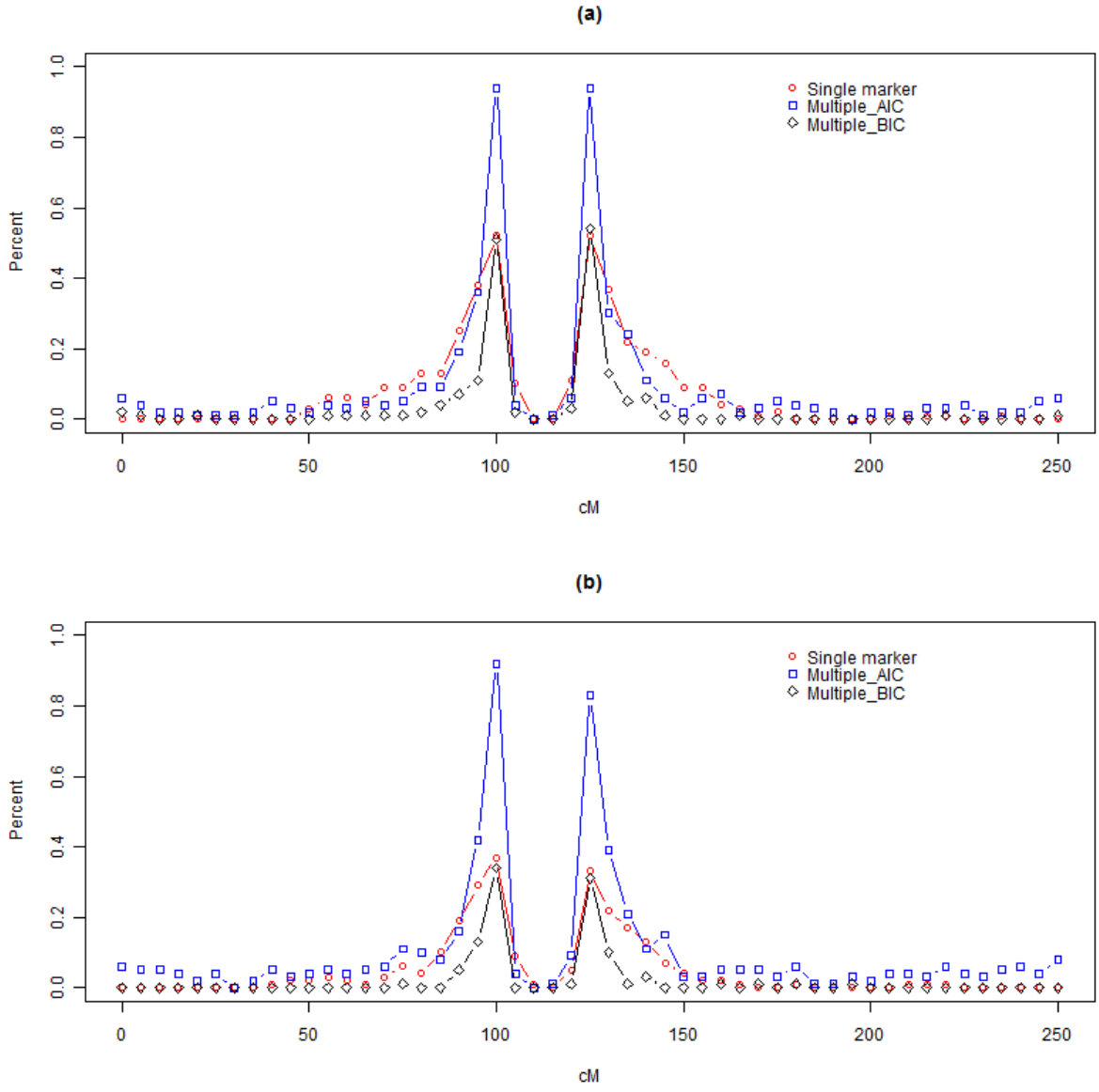

Figure 4.1: Proportion of selection for (a) Case 1 and (b) Case 2

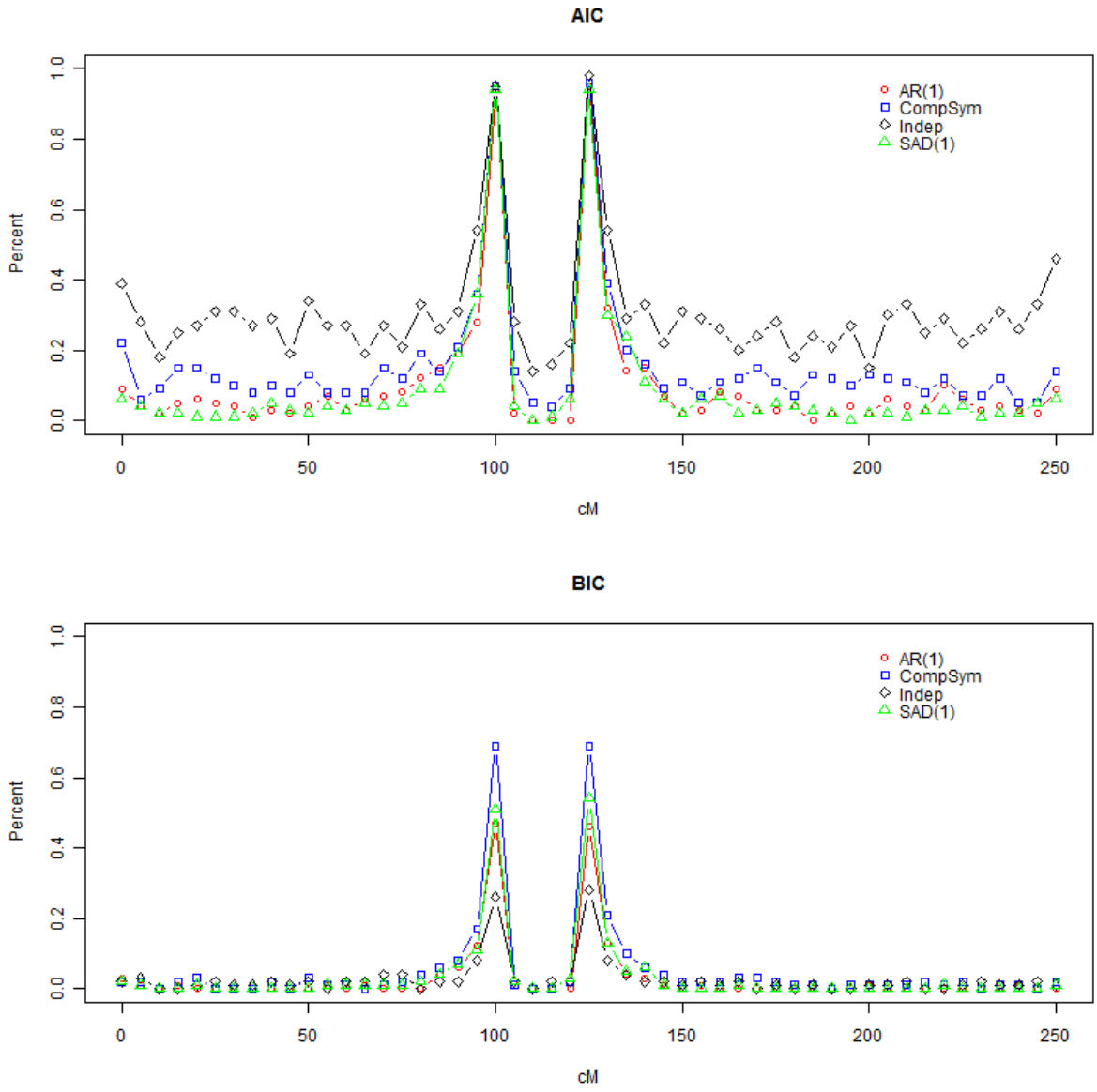

Figure 4.2: Proportion of selection for $\beta_1(t) = -\beta_2(t) = 1 + 3sin(\frac{\pi t}{30})$

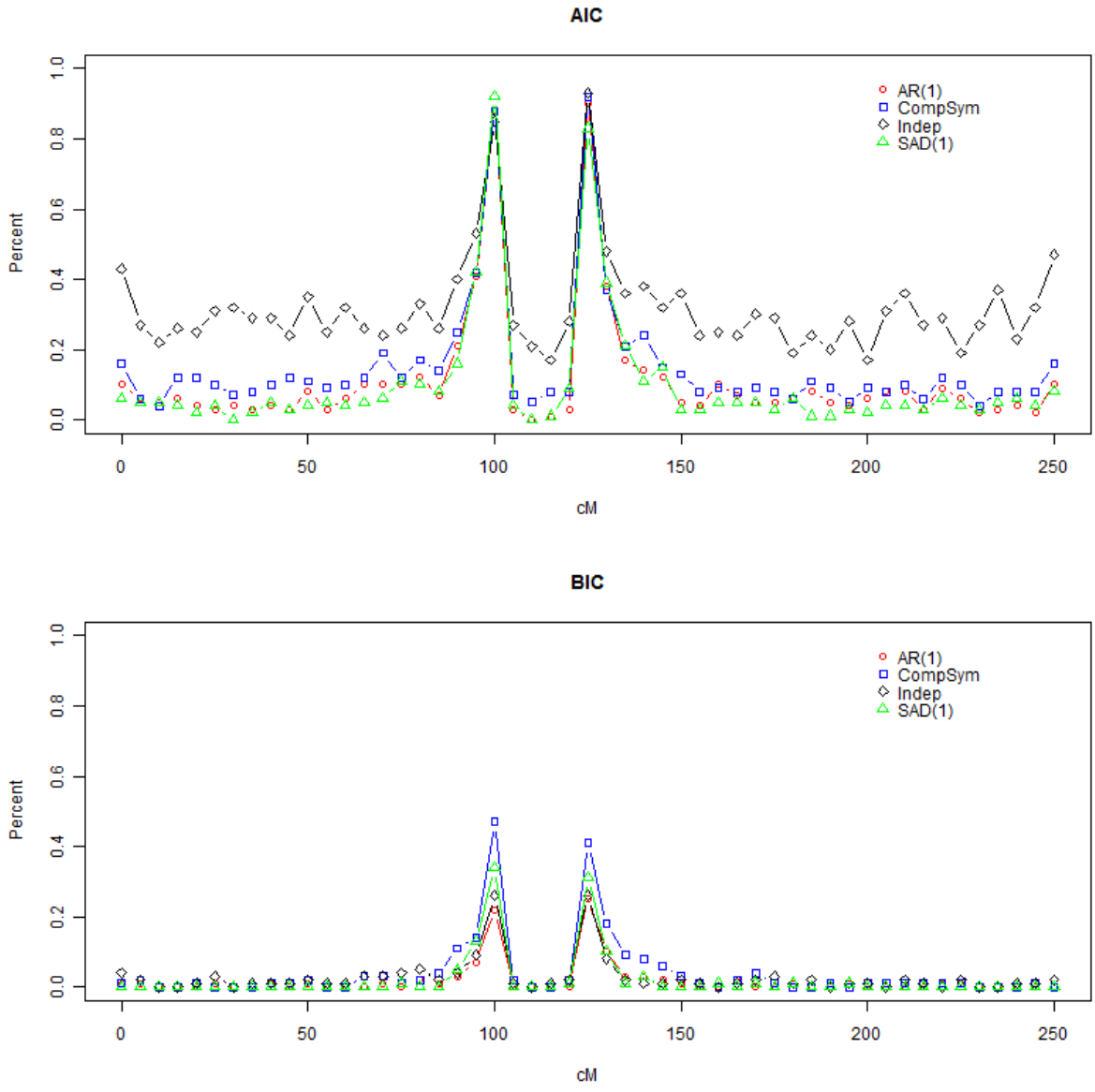

Figure 4.3: Proportion of selection for $\beta_1(t) = -\beta_2(t) = 1 + \frac{(30-t)^3}{5000}$

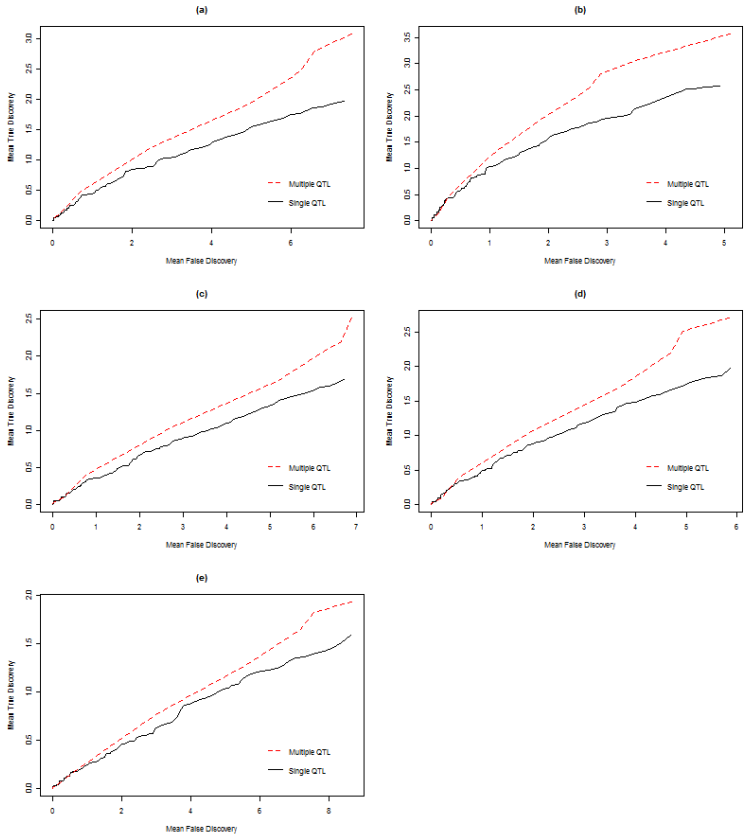

Figure 4.4: ROC-like plots comparing with $\sigma^2 = 120$

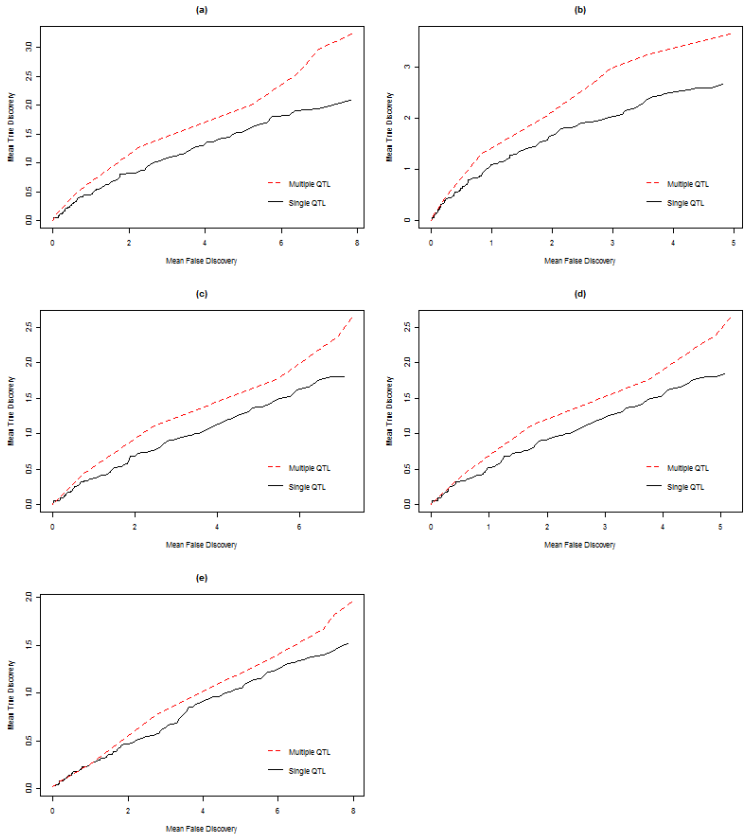

Figure 4.5: ROC-like plots comparing with $\sigma^2 = 80$

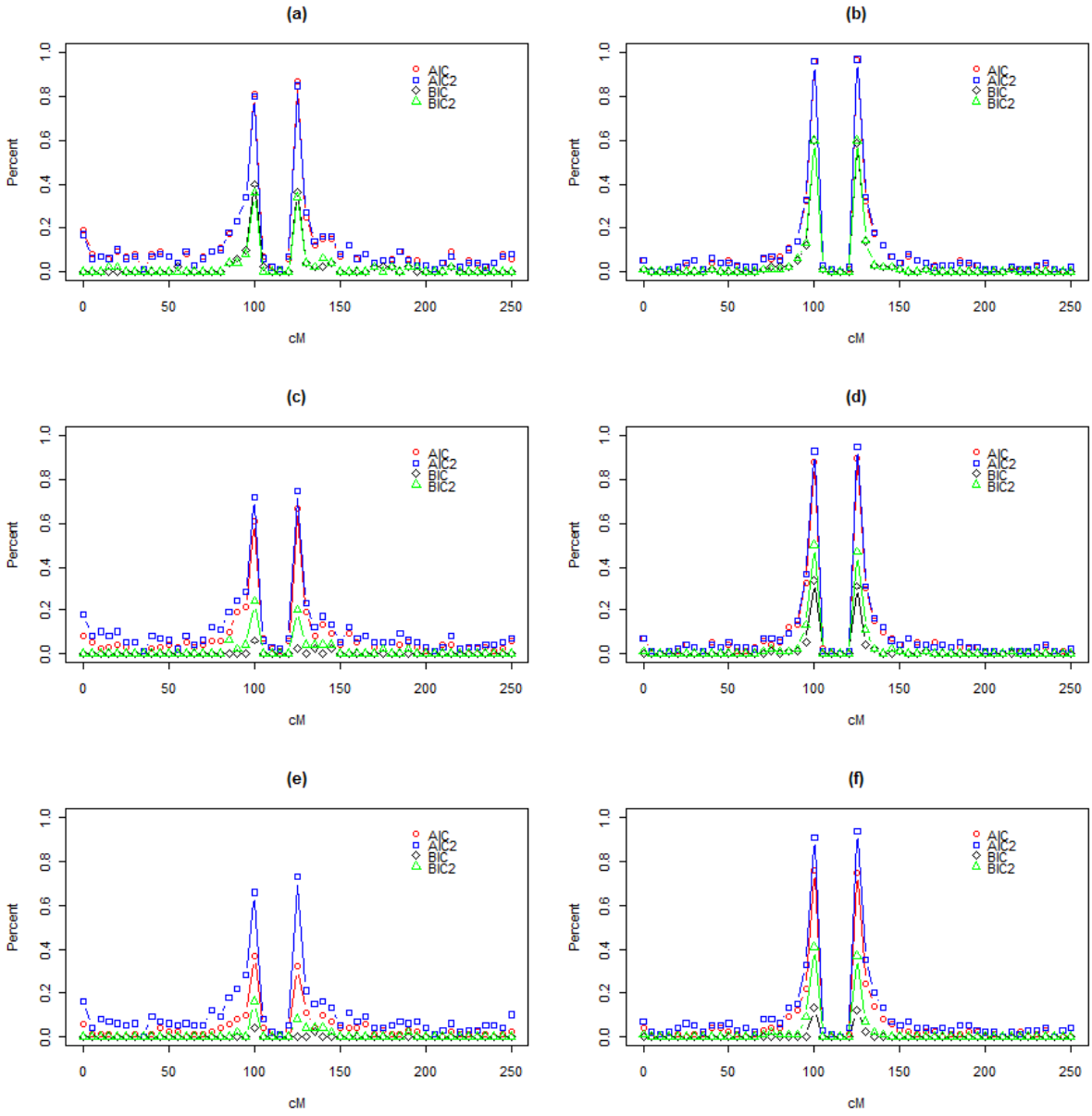

Figure 4.6: Proportion of selection using different criteria

References

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics* **22**, 203-217.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* **19**, 716-723.

Breiman, L. (1995). Better Subset Regression Using the Nonnegative Garrote. *Technometrics* **37**, 373-384.

Broman, K.W. and Speed, T.P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Methodological)* **64**, 641-656.

Cantoni, E., Flemming, J.M. and Ronchetti, E. (2005) Variable selection for marginal longitudinal generalized linear models. *Biometrics* **61**, 507-514.

Cheverud, J.M., Rutledge, J.J. and Atchley, W.R. (1983). Quantitative genetics of development: Genetic correlations among age-specific trait values and the evolution of ontogeny. *Evolution* **37**, 895-905.

Cleveland, W.S., Grosse, E. and Shyu, W.M. (1991). Local regression models. *In Statistical Models in S* (Chambers, J.M. and Hastie, T.J., eds), 309-376. Wadsworth & Brooks, Pacific Grove.

Craven, P. and Wahba, G. (1979). Smoothing Noisy Data With Spline Functions. *Numerische Mathematik* **31**, 377-403.

Darvasi, A. (1998). Experimental strategies for the genetic dissection of complex traits in animal models. *Nature Genetics* **18**, 19-24.

Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, K.Y. (2002). Analysis of Longitudinal Data. *Oxford University Press* Oxford, UK.

Donoho, D.L., and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** , 425455.

Dziak, J.J. and Li, R. (2006) An overview on variable selection for longitudinal data. In Hong, D. editor, *Quantitative Medical Data Analysis*. World Sciences Publisher, Singapore.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407-499.

Fan, J., and Gijbels, I. (1996). Local polynomial modelling and its applications. *Chapman and Hall* , London, UK.

Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348-1360.

Fan, J., and Zhang, J.T. (2000). Functional linear models for longitudinal data. *Journal of the Royal Statistical Society: Series B (Methodological)* **62**, 303-322.

Foulkes, W.D., Metcalfe, K., Sun, P., Hanna, W.M., Lynch, H.T., Ghadirian, P., Tung N., Olopade, O.I., Weber, B.L., Ivo, J.M., Olivotto, A., Begin, L.R. and Narod, S.A. (2004). Estrogen receptor status in BRCA1- and BRCA2-related breast cancer: the influence of age, grade, and histological type. *Clinical Cancer Research* **10**, 2029-2034.

Frank, I. and Friedman, J. (1993), A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-148.

Fu, W.J. (1998). Penalized regression: the bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**, 397-416.

Fu, W.J. (2003). Penalized Estimating Equations. *Biometrics* **59**, 126-132.

Fu, W.J. (2005). Nonlinear GCV and quasi-GCV for shrinkage models. *Journal of Statistical Planning and Inference* **131**, 333-347.

Gabriel, K.R. (1962). Ante-dependence analysis of an ordered set of variables. *The Annals of Mathematical Statistics* **33**, 201-212.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732.

Golub, G.H., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215-223.

Hastie, T.J. and Tibshirani, R.J. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)* **55**, 757-796.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer-Verlag, New York.

He, X.M. and Shi, P.D. (1998). Monotone B-spline smoothing. *Journal of the American Statistical Association* **93**, 643-650.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics* **12**, 69-82.

Hoover D.R., Rice, J.A., Wu, C.O. and Yang, L.P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809-822.

Huang, J., Ma, S., Xie, H. and Zhang, C. (2009). A group bridge approach for variable selection. *Biometrika* **96**, 339-355.

Huang, J., Ma, S. and Zhang, C.H. (2006). Adaptive lasso for sparse high-dimensional regression models. *Technical Report No. 374, Department of Statistics and Actuarial Science, University of Iowa* .

Huang, J.Z., Wu, C.O. and Zhou, L. (2002). Varying-Coefficient Models and Basis Function Approximation for the Analysis of Repeated Measurements. *Biometrika* **89**, 111-128.

Huang, J.Z., Wu, C.O. and Zhou, L. (2004). Polynomial Spline Estimation and Inference for Varying Coefficient Models With Longitudinal Data. *Statistica Sinica* **14**, 763-788.

Jansen, R.C. and Stam, P. (1994). High resolution mapping of quantitative traits into multiple loci via interval mapping. *Genetics* **136**, 1447-1455.

Jaffrezic, F., Thompson, F.R. and Hill, W.G. (2003). Structured antedependence models for genetic analysis of repeated measures on multiple quantitative traits. *Genetical Research* **82**, 55-65.

Kao, C.H. and Zeng, Z.B. (1997). General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics* **53**, 653-665.

Kao, C.H., Zeng, Z.B. and Teasdale, R.D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203-1216.

Kirkpatrick, M. and Heckman, N. (1989). A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *Journal of Mathematical Biology* **27**, 429-450.

Kirkpatrick, M., Lofsvold, D. and Bulmer, M. (1990). Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics* **152**, 979-993.

Knott, S.A. and Haley C.S. (1992). Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. *Genetical Research* **60**, 139-151.

Lander, E.S. and Botstein D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage results. *Genetics* **121**, 185-199.

Li, K.C. (1987). Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *Annals of Statistics* **15**, 958-975.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal Data Analysis using Generalized Linear models. *Biometrika* **73**, 13-22.

Lin, M. and Wu, R.L. (2006). A joint model for nonparametric functional mapping of longitudinal trajectories and time-to-events. *BMC Bioinformatics* **7**, 138.

Lin, Y. and Zhang, H.H. (2006). Component Selection and Smoothing in Smoothing Spline Analysis of Variance ModelsCOSSO. *Annals of Statistics* **34**, 2272-2297.

Ma, C., Casella, G. and Wu, R.L. (2002). Functional mapping of quantitative trait loci underlying the character process: A theoretical framework. *Genetics* **161**, 1751-1762.

Mallows, C.I. (1973). Some comments on C_p. *Technometrics* **15**, 661-676.

Manichaikul, A., Moon, J.Y., Sen, S., Yandell, B.S. and Broman, K.W. (2009). A Model Selection Approach for the Identification of Quantitative Trait Loci in Experimental Crosses, Allowing Epistasis. *Genetics* **181**, 1077-1086.

Martinez, O. and Curnow, R.N. (1992). Estimation the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85**, 480-488.

Meier, L., van de Geer, S. and Buhlmann P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Methodological)* **70**, 53-71.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis* **52**, 374-393.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* Chapman and Hall: London.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear and Mixed Models.* Wiley-Interscience, New York.

Nelder, J.A., and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal* **7**, 308-313.

Nunez-Anton, V. (1997). Longitudinal data analysis: Non-stationary error structures and antedependent models. *Applied Stochastic Models and Data Analysis* **13**, 279-287.

Nunez-Anton, V. and Zimmerman, D.L. (2000). Modeling nonstationary longitudinal data. *Biometrics* **56**, 699-705.

Pan, W. (2001) Akaikes Information Criterion in generalized estimating equations. *Biometrics* **57**, 120-125.

Pauler, D.K. (1998) The Schwarz criterion and related methods for normal linear models. *Biometrika* , **85**, 13-27.

Picard, R.R. and Cook, R.D. (1984) Cross-validation of regression models. *Journal of the American Statistical Association* **79**, 575-583.

Pittman, J. (2002). Adaptive splines and genetic algorithms. *Journal of Computational and Graphical Statistics* **11**, 615-638.

Pletcher, S.D. and Geyer, C.J. (1999) The genetic analysis of age-dependent traits: modeling the character process. *Genetics* **153**, 825-835.

Plomin, R., McClearn, G.E., Gora-Maslak, G. and Neiderhiser, J.M. (1991). An RI QTL cooperative data bank for recombinant inbred quantitative trait loci analyses. *Behavior Genetics* **21**, 97-98.

Schaeffer L.R. (2004). Application of random regression models in animal breeding. *Livestock Production Science* **86**, 35-45.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics* **19**, 461-464.

Shao,J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**, 486-494.

Shao,J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica* **7**, 221-264.

Shaw, R.G. (1987). Maximum-likelihood approaches applied to quantitative genetics of natural populations. *Evolution* **41**, 812-826.

Sun, W., Ibrahim, J.G., and Zou, F. (2010). Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. *Genetics* **185**, 349-359.

Threadgill, D.W., Hunter, K.W. and Williams, R.W. (2002). Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mammalian Genome* **13**, 175-178.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267288.

Tibshirani, R., Saunders, M., Rosset, S. Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **67**, 91-108.

Vaughn, T.T., Pletscher, L.S., Peripato, A, King-Ellison K., Adams, E., Erikson, C. and Cheverud, J.M. (1999). Mapping quantitative trait loci for murine growth: a closer look at genetic architecture. Genetical Research **74**, 313-322.

Wang, L., Chen, G., and Li, H. (2007). Group SCAD Regression Analysis for Microarray Time Course Gene Expression Data. *Bioinformatics* **23**, 1486-1494.

Wang, L. Li, H. and Huang, J.Z. (2008). Variable Selection in Nonparametric Varying-Coefficient Models for Analysis of Repeated Measurements. *Journal of the American Statistical Association* **103**, 1556-1569.

Wolfinger, R.D., Tobias, R.D., and Sall, J. (1994). Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM Journal on Scientific Computing* **15**, 1294-1310.

Wright F.A. and Kong A. (1997). Linkage mapping in experimental crosses: the robustness of single-gene method. *Genetics* **146**, 417-425.

Wu, R.L., Ma, C., Littell, R. and Casella, G. (2002). A statistical model for the genetic origin of allometric scaling laws in biology. *Journal of Theoretical Biology* **217**, 275-287.

Wu, R.L., Ma, C., Lin, M. and Casella, G. (2004). A general framework for analyzing the genetic architecture of developmental characteristics. *Genetics* **166**, 1541-1551.

Yang, J., Wu, R.L. and Casella, G. (2009). Nonparametric Modeling of Longitudinal Covariance Structure in Functional Mapping of Quantitative Trait Loci. *Biometrics* **65**, 30-39.

Yi, N., and Xu, S. (2000). Bayesian mapping of quantitative trait loci for complex binary traits. *Genetics* **155**, 1391-1403.

Yuan, M. and Lin, Y. (2006). Model Selection and Estimation in Regression With Grouped Variables. *Journal of the Royal Statistical Society: Series B (Methodological)* **68**, 49-67.

Zeger, S.L. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.

Zeng, Z.B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences USA* **90**, 10972-10976.

Zeng, Z.B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457-1468.

Zeng, Z.B., Liu, J., Stam, L.F., Kao, C.H., Mercer, J.M., et al. (2000). Genetic architecture of a morphological shape difference between two Drosophila species. *Genetics* **154**, 299-310.

Zhang, H. and Lin, Y. (2006). Component Selection and Smoothing for Nonparametric Regression in Exponential Families. *Statistica Sinica* **16**, 1021-1042.

Zhang, H. and Zhong, X. (2006). Linkage analysis of longitudinal data and design consideration. *BMC Genetics* **7**, 37.

Zhao, P. and Yu, B. (2007). On model selection consistency of lasso. *Journal of Machine Learning Research* **7**, 2541-2567.

Zhao, W., Hou, W., Littell, R.C. and Wu, R.L. (2005). Structured Antedependence Models for Functional Mapping of Multiple Longitudinal Traits. *Statistical Applications in Genetics and Molecular Biology* **4**, 33.

Zhao, W., Ma, C., Cheverud, J.M. and Wu, R.L. (2004). A unifying statistical model for QTL mapping of genotype $\times$ sex interaction for developmental trajectories. *Physiological Genomics* **19**, 218-227.

Zou, F., Gelfond, J.L., Airey, D.C., Lu, L., Manly, K.F., Williams, R.W. and Threadgill, D.W. (2005). Quantitative trait locus analysis using recombinant inbred intercrosses (RIX): theoretical and empirical considerations. *Genetics* **170**, 1299-1311.

Zou, H. (2006). The Adaptive LASSO and Its Oracle Properties. *Journal of the American Statistical Association* **101**, 1418-1429.

Zou, H. and Hastie T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Methodological)* **67**, 301-320.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "Degrees of Freedom" of the Lasso. *Annals of Statistics* **35**, 2173-2192.

Zou, H. and Li, R. (2008). One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models. *Annals of Statistics* **36**, 1509-1533.